LEABHARLANN CHOLÁISTE NA TRÍONÓIDE, BAILE ÁTHA CLIATH | TRINITY COLLEGE LIBRARY DUBLIN
Ollscoil Átha Cliath | The University of Dublin

Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Vocal correlates of affective states

Irena Yanushevskaya

Centre for Language and Communication Studies
School of Linguistic, Speech and Communication Sciences
Trinity College Dublin

Submitted for the degree of PhD

2010

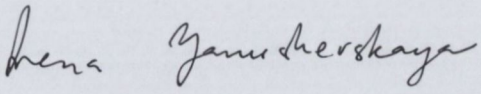
TRINITY COLLEGE
27 AUG 2010
LIBRARY DUBLIN
THESIS
8931

A MI MADRE

Y A LA MEMORIA DE MI PADRE

Declaration

I declare that this thesis is my own work unless stated otherwise in the text. This thesis has not been submitted for a degree at any other university. I agree that the Library of Trinity College Dublin may lend or copy the thesis upon request.



Irena Yanushevskaya

Summary

This thesis explores the mapping of voice quality to affect. Three studies have been conducted.

The principal study of this thesis is a cross-language study, aimed at clarifying how variations in voice quality (in synthesised stimuli) can evoke different affective colouring for subjects from four different language backgrounds: Hiberno-English, Russian, Spanish and Japanese. This study furthermore addresses, by including major f_0 differences in some stimuli, some aspects of the role of f_0 in affect cueing.

It was of interest to ascertain (1) whether voice quality cues become more effective when major f_0 perturbations are included, (2) what is the contribution of dynamic f_0 contour when voice quality is kept modal, (3) whether and to what extent the stimuli incorporating these voice quality and f_0 variations impart similar/different affective colouring to the listeners from different language groups. In the listening tests, the listeners were presented with three types of stimuli: (1) 'VQ only' stimuli with distinct voice quality variations (whispery, breathy, lax-creaky, tense and modal); (2) 'f_0 only' stimuli in all of which modal voice quality was combined with affect-related f_0 variations (boredom, sadness, joy, indignation, fear, and neutral); (3) 'VQ + f_0' stimuli in which distinct non-modal voice qualities were combined with certain affect-related f_0 contours. The strength of affective content of each stimulus was assessed for a number of affective labels: *apologetic-indignant, bored-interested, intimate-formal, relaxed-stressed, sad-happy* and *scared-fearless*

The results of the study showed broad cross-language similarities in the way voice maps to affect, as well as some major cross-language differences.

For all languages tested, all non-modal 'VQ only' stimuli get associated with a number of affects. In most cases, there is no one-to-one voice-to-affect mapping. Similarly, the same affect can be cued by more than one voice quality. Overall, for the range of stimuli included in this experiment, those which incorporate distinct voice qualities ('VQ only' or 'VQ + f_0' series) emerged as being the most effective in signalling affect. The 'f_0 only' stimuli appear to be relatively less effective. This holds for all the languages tested, and there is considerable cross-language agreement in the way these stimuli signal the affects *sad, scared, indignant, apologetic, interested, bored,* and *fearless*.

Prominent cross-language differences emerged for the subtests *intimate-formal* and *relaxed-stressed*. *Intimate* was associated for Spanish and Japanese subjects with tense/modal voice combined with f_0 'indignation', a stimulus which was associated with *formal* for the Hiberno-English and Russian subjects. *Intimate* was cued by lax-creaky voice or whispery voice for Hiberno-English and Russian but not at all for the Spanish and Japanese listeners. Furthermore, in these subtests, there was a conspicuous gap in the cueing of *formal*, *stressed* and *relaxed* with any of the stimuli presented here for the Japanese listeners, but not for the speakers of the European languages tested. These large differences in voice to affect mapping are likely to reflect underlying

cross-cultural differences as they involve attitudes and interpersonal stances, which are more likely to reflect culture-shaped learned behaviour. Other factors, such the differences in semantic fields of the affective labels used in the tests for the different languages, or the differences in neutral voice (baseline) in these languages might have contributed to the results.

Further more subtle cross-language differences emerged that suggest that languages may differ in the relative sensitivity to voice quality or f_0 cues. E.g., there are indications that Japanese subjects may be more sensitive to f_0 cues where the European group of languages rely more heavily on voice quality cues.

Different voice quality stimuli are characterised by different loudness. Two related 'loudness experiments' were conducted to explore to what extent these loudness differences might have contributed to the affective colouring these voice quality stimuli impart. Overall, the results suggest that loudness variation on its own is rather ineffective for affect cueing while variation in voice quality is essential to the expression of affect. In Loudness experiment 1, stimuli incorporating voice quality variations consistently obtained higher ratings compared to the stimuli varying in loudness but with the same modal voice quality. The results of Loudness experiment 2 showed that equalising loudness of voice quality stimuli by setting their perceived loudness to match that of the modal voice does not have significant impact on their affective ratings: the voice quality stimuli appear capable of affect cueing for the majority of affects tested. However, loudness does contribute in affect signalling: when combined with tense-modal voice qualities, increased loudness can enhance signalling of high activation states such as *indignant, stressed* and *happy*.

The production based study involved a detailed voice source analysis on an utterance produced by a non-actor portraying the following affective states: *bored, sad, happy, surprised, angry* as well as *neutral*. To characterise the voice source variation in these portrayals, the voice source parameters were obtained through the inverse filtering of the utterance and the subsequent LF model matching of the resulting differentiated glottal flow waveform. As the speech sample was not representative (portrayed affect; a single speaker; one utterance for each of the portrayed affects), the results are not presented to make any kind of definitive claims about how affects differ. Rather, this study was intended as an illustration of how a detailed parameter-rich analysis can proceed and to explore the kinds of dimensions that may be important in affect differentiation. The study served to illustrate (1) that a parameter rich description of voice source variation, rather than global one-parameter characterisation, is more likely to capture the complexity of the differences that may occur; (2) that there is a very considerable within-utterance dynamic variation; (3) that important aspects of dynamic variation seem to be linked to certain prosodically important points and that the study of affect-related source variation needs to be conducted in parallel with the study of the underlying prosody of the voice.

Acknowledgements

I wish to express my sincere gratitude to my supervisors Prof. Ailbhe Ní Chasaide and Dr. Christer Gobl for introducing the fascinating subject of voice quality to me, for sharing their ideas in many productive discussions, as well as for the unparalleled support throughout my work on this thesis. I am also thankful to the examiners of my thesis Prof. Roddy Cowie and Prof. Nick Campbell for their patience and for their very helpful and detailed critical comments.

I thank all those who kindly volunteered to assist me in data collection for the cross-cultural perception experiments in Ireland, Japan, Spain and Russia as well as for other perception experiments in this thesis. I owe special thanks to Áine Ní Bhriain, Brian Ó Raghallaigh, Martha Dalton and Michelle Tooher (Dublin); Mika Ito and Ken-Ichi Sakakibara (Tokyo); Ricard Franch, Pau Cuevas and Alfons Claver (Barcelona); Elena A. Shamina and Elena A. Varzhavitina (St. Petersburg). I would also like to use this opportunity to thank all the participants in the listening tests – your contribution has been crucial.

I am very grateful to those who generously shared their time and expert knowledge to help me resolve some of the problems I encountered during my work on this thesis: Mika Ito, Denise O'Leary, Áine Ní Bhriain, Brian Ó Raghallaigh, Michelle Tooher, Myra O'Regan, Eamonn Mullins and Cathal Walsh.

My special thanks go to my colleagues at the Speech and Language Therapy Department at the NUI Galway, for their support, and in particular to Rena Lyons. Thanks also to all the good folks at the Phonetics and Speech Laboratory at TCD.

I am most grateful to all my friends, old and new, for their support and good humour, especially to Anya & Julia, E.A., Áine, Susana, Laura, Brian, John, and another John, Christian, and Greg.

My deepest gratitude is to my parents for their love, patience and understanding, and for being there for me.

I gratefully acknowledge the financial support that I received during my work on this thesis. It was funded by Trinity College Dublin Studentship, EU 6th Framework Network of Excellence on Human-Machine Interaction (HUMAINE) and Hitachi Ltd.

Table of contents

SUMMARY I

ACKNOWLEDGEMENTS III

TABLE OF CONTENTS V

1 CHAPTER ONE INTRODUCTION 1

1.1 Introduction 1
1.2 Aims 1
1.3 Inherent difficulties in investigating voice quality correlates of affect 2
1.3.1 Obtaining reliable robust measures of voice quality 2
1.3.2 Eliciting affectively coloured speech data 3
1.3.3 Classifying and labelling emotion/affect 3
1.3.4 Classifying and labelling voice qualities 3
1.4 Background to the cross-language study 4
1.5 Interaction of voice quality, f_0 and intensity: further background to the perception study 5
1.6 The production-based study 6
1.7 Research questions 7
1.8 Structure 8

2 CHAPTER TWO VOICE AND EMOTION 9

2.1 Voice 9
2.1.1 Introduction 9
2.1.2 Auditory description 11
2.1.3 Measuring the acoustic correlates of voice quality 16
2.1.4 Vocal expression of emotion 27
2.2 Emotion 30
2.2.1 Terminology and definitions 30
2.2.2 Theories of emotion 34
2.3 Vocal expression of emotion 42
2.3.1 Collecting emotional speech data 42
2.3.2 Perceptual approaches 46
2.3.3 Vocal expression of emotion: acoustic measures 52
2.4 Universality and culture specificity of vocal expression of emotion 68
2.5 Summary 77

3 CHAPTER THREE CROSS-LANGUAGE STUDY: METHODOLOGY 79

3.1 Introduction 79
3.2 Materials: the synthesised stimuli 82
3.2.1 'VQ only' stimuli: modal voice 82
3.2.2 'VQ only' stimuli: non-modal voice qualities 83
3.2.3 'f_0 only' stimuli 85
3.2.4 'VQ + f_0' stimuli 89
3.3 Method: the perception tests 92
3.3.1 Scales 92
3.3.2 Affective labels 93
3.3.3 Participants 101
3.4 Scope and limitations of the study 105
3.5 Research questions 107
3.6 Statistical analysis 107

3.6.1 Design 107
3.6.2 Interrater agreement 108

4 CHAPTER FOUR VOICE-TO-AFFECT MAPPING: CROSS-LANGUAGE STUDY 111

4.1 Introduction 111
4.2 Apologetic-Indignant 115
4.2.1 The statistical analyses 115
4.2.2 Apologetic 116
4.2.3 Indignant 119
4.3 Bored-Interested 121
4.3.1 The statistical analyses 121
4.3.2 Bored 122
4.3.3 Interested 125
4.4 Intimate-Formal 128
4.4.1 The statistical analyses 128
4.4.2 Intimate 129
4.4.3 Formal 134
4.5 Relaxed-Stressed 137
4.5.1 The statistical analyses 137
4.5.2 Relaxed 138
4.5.3 Stressed 141
4.6 Sad-Happy 144
4.6.1 The statistical analyses 144
4.6.2 Sad 145
4.6.3 Happy 147
4.7 Scared-Fearless 149
4.7.1 The statistical analyses 149
4.7.2 Scared 149
4.7.3 Fearless 152
4.8 Summary and conclusions 154

5 CHAPTER FIVE LOUDNESS AND VOICE QUALITY 162

5.1 Introduction 162
5.2 Loudness: theoretical background 162
5.2.1 Definition of loudness 162
5.2.2 Measuring loudness 163
5.2.3 Factors influencing loudness perception: properties of the signal 168
5.2.4 Factors influencing loudness perception: the listener 172
5.2.5 Physiological correlates of loudness 173
5.2.6 Loudness in speech studies 174
5.2.7 Loudness in vocal expression of emotion 177
5.3 Loudness experiment 1 179
5.3.1 Introduction 179
5.3.2 Material: synthesised stimuli 179
5.3.3 Preliminary listening test: obtaining loudness matches 180
5.3.4 Listening tests 183
5.3.5 Results and discussion 185
5.3.6 Conclusions 192
5.4 Loudness experiment 2 193
5.4.1 Introduction 193
5.4.2 Material: synthesised stimuli 193
5.4.3 Listening tests 195
5.4.4 Results 196

5.4.5 Summary and discussion 200
5.4.6 Conclusions 203
5.5 Summary 204

6 CHAPTER SIX VOICE SOURCE PARAMETERS IN AFFECT PORTRAYALS: AN ILLUSTRATIVE ANALYSIS 206

6.1 Introduction 206
6.2 Material and method 208
6.2.1 Speech sample 208
6.2.2 Listening test 209
6.2.3 Data analysis 210
6.2.4 Statistical analysis and data processing 212
6.3 Results and discussion 214
6.4 Conclusions 223

7 CHAPTER SEVEN CONCLUSIONS 225

7.1 Introduction 225
7.2 Main findings and suggestions for future work 225
7.2.1 Cross-language study 225
7.2.2 Voice quality and loudness 228
7.2.3 Voice source parameters in affect portrayals: an illustrative analysis 229

REFERENCES 233

1 APPENDIX 1 265

2 APPENDIX 2 275

3 APPENDIX 3 289

4 APPENDIX 4 303

5 APPENDIX 5 307

6 APPENDIX 6 311

7 APPENDIX 7 331

8 APPENDIX 8 345

9 APPENDIX 9 349

10 APPENDIX 10 363

11 APPENDIX 11 366

12 APPENDIX 12 367

13 APPENDIX 13 375

14 APPENDIX 14 379

15 APPENDIX 15 387

16 APPENDIX 16 390

17 APPENDIX 17 391

1 Chapter one Introduction

1.1 Introduction

The human voice is a fascinating instrument, and its role in human interaction is fundamental. We are left in awe by the art of operatic singing and the subtle nuances in interpretation of the same aria by different performers. In everyday life, the voice of the people we know tells us whether they are sad and worried or happy and excited even if the verbal message contains no affective information and is transmitted through such a distortion-introducing channel as a mobile phone. In interpersonal communication, we skilfully manipulate the tone of voice to signal (or mask) our true emotions or to simulate feelings and attitudes that we consider beneficial for our well-being. Voice and the way we use it form an important part of our identity. Disorders resulting in the loss of one's habitual tone of voice may sometimes entail serious consequences for one's self-image and self-presentation.

The last thirty years have seen a growing body of research in the field of vocal expression and communication of emotion (e.g., a review in Scherer, 2003). However, still relatively little is known about how voice source varies in emotion signalling. Most of the data on vocal expression have been collected for methodologically less problematic characteristics of the tone of voice, such as f_0, intensity and timing. It is therefore important to broaden the knowledge related to the way voice quality (which depends in a narrow sense on the voice source characteristics related primarily to the mode of vibration of the vocal folds) is used in affect signalling as well as the way it is combined with f_0 variations. Another aspect of speech communication on which relatively little empirical data exist concerns universality and language- or culture-specific use of the vocal cues in signalling emotion. The thesis aims at throwing light on this problem.

1.2 Aims

The thesis explores the mapping of voice quality to affect. There are three studies on this theme.

1. **The cross-language study**. The principal study of this thesis is a perceptual study, aimed at clarifying how variations in voice quality (in synthesised stimuli) can evoke different affective colouring for subjects

from four different language/cultural backgrounds. This study furthermore addresses, by including major f_0 differences in some stimuli, some aspects of the role of f_0 in affect cueing, particularly for the signalling of strong emotions.

2. **The loudness study**. An ancillary study, involving two perceptual experiments, considers the role of loudness variation (along with voice quality variation) in affect cueing.

3. **The production based study**. Detailed voice source analysis is carried out in an utterance produced by a non-actor subject in ways that were intended to portray the following affective states: *bored, sad, happy, surprised, angry* as well as *neutral.*

1.3 Inherent difficulties in investigating voice quality correlates of affect

The importance of voice quality in the vocal expression and communication of emotion is widely acknowledged. However, due to a number of methodological difficulties and the lack of empirical data, still relatively little is known about voice quality in emotion signalling. In particular, the following problems arise.

1.3.1 Obtaining reliable robust measures of voice quality

The analytical task of obtaining voice source data is not straightforward. Automatic methods for inverse filtering, the main method for estimating voice source signal, are likely to introduce error and are therefore unreliable, whereas more reliable manual interactive methods are time consuming and allow analysis of only small amounts of data (Ní Chasaide & Gobl, 1997). Moreover, the recording conditions have to be very stringent for accurate inverse filtering.

The voice source varies considerably with the segmental and prosodic context of the utterance, even in a sentence where there is a neutral (modal) voice quality and no discernible affective colouring. Although there has been some discussion of this aspect of voice source variation (Gobl, 2003; Gobl & Ní Chasaide, 1999b; Ní Chasaide & Gobl, 1997, 2004a), it remains largely undescribed.

Thus there is the difficulty of interpreting raw voice source data. Unless one can separate out these baseline effects of cross-speaker variation as well as segment- and prosody-sensitive variation, it is difficult to be sure that particular estimates of the voice source, even if accurately measured, are in fact related to the affective content of a

particular utterance and not to the segmental/prosodic environment from which the measures are taken, or to characteristics of a speaker's voice.

1.3.2 Eliciting affectively coloured speech data

For the reasons outlined above it would be desirable to obtain voice source data where the segmental and prosodic content is controlled, that is the same utterance, produced with different affects. This approach has been widely used (Banse & Scherer, 1996; Scherer, 2003), and the simulated emotions have been produced by actors or non-actors.

The criticism of such an approach is that the emotions portrayed are not genuine, and that such vocal expression of emotion may not compare truly to genuinely arising spontaneous emotional vocalisations. Using induction techniques to elicit particular affective states (such as stress) in the subjects may partially alleviate this, but can be problematic due to frequently arising ethical issues.

A different approach is to gather unscripted naturalistic data, whether from radio and TV shows or in a specifically recorded speech corpus. Analysing such data raises many problems, including those just mentioned such as catering for the uncontrolled effects of the segmental and prosodic context, the recording conditions, the reliability of the voice analysis, the labelling of the expressed affective states etc.

1.3.3 Classifying and labelling emotion/affect

There is a vast literature on the theory of emotion, and on the inherent difficulties in labelling emotions with a reliable frame of reference. What one researcher refers to as 'angry' is not necessarily the same entity as that referred to as 'angry' by another researcher. [How do we know that what researcher A refers to as angry is the same entity as that referred to as angry by researcher B?] There is an ongoing debate as to whether emotions are discrete entities or whether they might be best analysed in terms of affective dimensions, such as activation, valence, power etc.

The labelling issue is even more likely to be problematic if cross-language data is considered, given that the terms defining affective states may not be fully semantically equivalent in the different languages.

1.3.4 Classifying and labelling voice qualities

Similar to the classification and labelling of emotion/affect, there is a major difficulty in pinning down what precisely is meant by voice quality terms. Impressionistic auditory labels such as 'bright voice', 'husky voice' or 'guttural voice' are frequently used

but rarely formally defined, and they are likely to be associated with a different auditory percept by different researchers. Conversely, different terms, e.g., 'harsh', 'rough', 'hoarse' can be used to describe essentially the same voice quality. As pointed out in Gobl and Ní Chasaide (2003b), it is not always clear whether the differences/similarities in vocal expression of affect result from the genuine voice quality related phenomena or simply from the diversity of the use of voice quality defining labels.

Laver's phonetic description of voice quality (Laver, 1980) offers a well defined classification system of phonation types which has been crucial for establishing a consistent descriptive framework for those working in the area of voice quality.

Many of these issues are discussed in Chapter 2, where the relevant literature is reviewed.

1.4 Background to the cross-language study

The principal study of this thesis, the cross-language perception study, follows an approach which has been used in Trinity College Dublin in recent years (Gobl, Bennett, & Ní Chasaide, 2002; Gobl & Ní Chasaide, 2003b; Ní Chasaide & Gobl, 2005). Rather than record and analyse affectively coloured speech, the approach attempts at eliciting listeners' affective attributions to a range of different voice qualities. This involves having subjects listen to an utterance synthesised with a variety of voice qualities and having them rate whether and to what extent these stimuli impart affective overtones.

The voice quality stimuli were essentially the same as those used in earlier studies (Gobl & Ní Chasaide, 2003b). They were based on prior analyses (Gobl & Ní Chasaide, 1992; Ní Chasaide & Gobl, 1997) as well as by the broader literature on the acoustic characteristics. The studies carried out in Trinity College Dublin all used subjects who were speakers of Hiberno-English. The results of these studies showed rather consistent affective response being associated with particular voice qualities.

One of the main questions addressed in the cross-language study (Chapter 4) is whether the mapping of voice quality to affect is universal or language specific. This thesis extends earlier research in that subjects from four different language/culture backgrounds were tested: subjects from Russia, Spain and Japan were tested along with subjects from Ireland, speakers of Hiberno-English.

1.5 Interaction of voice quality, f_0 and intensity: further background to the perception study

Voice quality is often referred to as tone of voice. It relates at the acoustic level to the complex dimensions of the voice source, which at the physiological level are determined primarily by the laryngeal tension settings and muscular adjustments, along with variations in the respiratory forces and airflow rates through the glottis, e.g., Titze (1994). It is important to note however that voice quality is a perceptual phenomenon: although there are typical acoustic correlates of a particular voice quality, if we find a particular voice source parameter value in our data, this doesn't unequivocally mean we can identify the auditory voice quality.

In the phonetic literature (see Laver, 1980) voice qualities, such as tense voice, breathy voice or creaky voice are seen as global, holistic percepts that include some differences in pitch and loudness. In a great deal of the literature on emotion however, voice quality, intensity and f_0 are treated as separate entities, with the (typically unspoken) assumption that they are under separate control and operate as relatively independent cues in perception.

To some extent, this is of course true. Intensity and f_0 are potentially under independent control: it is possible to produce, say, breathy voice at different levels of vocal effort, and with different pitches, and undoubtedly speakers do this. However, differences in loudness and, to some extent pitch, are typically part and parcel of voice quality variation.

In the principal study of this thesis, the cross-language perception study, a range of voice quality stimuli were used which included, as an inherent part of these voice qualities, differences in intensity and in f_0.

The differences in f_0 were, for the most part, rather slight (see details in the methodology descriptions in Chapter 3). They did not include the large perturbations, e.g., the major increases in f_0 level and dynamics reported for strong emotions such as joy or anger by many researchers (see review of the literature in Chapter 2).

For that reason, while the main focus of the cross-language study is to investigate how subjects from different language/culture backgrounds map voice quality to emotion, it was also felt useful to include stimuli where large f_0 perturbations were combined with certain voice qualities, to see whether such manipulations (such combined stimuli) would yield higher ratings for these strong emotions. Furthermore, as a control, stimuli with the same highly differing f_0 patters, but with modal voice quality were included to rule out the possibility that the f_0 pattern alone might be responsible for any demonstrable effect.

As for intensity, the stimuli used in the cross-language study inevitably had different intensity levels. While this was intended and unavoidable when one varies the glottal pulse, the question did arise as to whether the high affective ratings associated by subjects with certain voice qualities might conceivably be the result of the stimuli's inherent differences in perceived loudness. To check this out, two further perception experiments were designed, and reported in Chapter 5. The first simply elicited whether differences in loudness per se could suffice to elicit the affective responses of the cross-language study. This was tested by varying the loudness of the modal voice quality stimulus, so that it matched the loudness of the other non-modal voice qualities, and testing listeners on both these loudness varying modal voice stimuli and the original voice quality stimuli of the earlier experiment. The second experiment looked more closely at the contribution of loudness differences to the affective ratings of different voice qualities. In this experiment, the loudness of the different voice qualities themselves was manipulated, and the perceptual consequences in terms of affective ratings checked.

1.6 The production-based study

This was a short descriptive study, aimed at providing very detailed voice source information on a fragment of production data recorded using a semi-anechoic recoding studio and phase linear recording equipment. The data involved a single utterance produced by a non-actor with the intension of portraying *bored, sad, happy, surprised, angry*, as well as the *neutral* state. These data do not make any claims about ecological validity of the affects included, and would only describe a possible way that these can be portrayed (e.g., by a narrator reading a story).

The intention was (i) to illustrate the extent to which these renditions can be characterised by different combinations of voice source parameter settings, and (ii) to illustrate some of the dynamic aspects of within utterance variation.

There was a further aim, though not one that is pursued in this study: to demonstrate whether resynthesis of these parameter settings could be used to generate synthetic utterances with audibly similar affective colouring. These latter aims were not pursued in this study (but are reported in Tooher et al. (2008)) and will not be discussed further here.

The analysis involved manual interactive inverse filtering of the utterance, on a pulse-by-pulse basis, to obtain glottal source signal estimates. Following this, as a way of extracting voice source parameters to describe the data, the LF model of the glottal source

(Fant, Liljencrants, & Lin, 1985) was matched to the voice source signal. The analysis methods are outlined in detail in Chapter 2, with further details provided in Chapter 6.

1.7 Research questions

At this point it is useful to state the specific research questions addressed in the different studies.

The cross-language perception study

- How does voice quality map to affect? Does this mapping support traditional description found in the literature?
- To what extent are extreme f_0 perturbations required for the signalling of strong emotions? Does the addition of a dynamic f_0 contour enhance affective rating of a particular voice quality stimulus?
- To what extent are the affective consequences (i.e. the affective ratings yielded when combining major f_0 perturbations with voice qualities) achieved anyway by simply the f_0 perturbations alone? To what extent can the f_0 contour (alone, i.e. with modal voice) induce the same affective rating?
- Do languages associate voice quality (and f_0 contours) with affect in a similar way, or do distinct language-specific trends emerge?

The loudness/voice quality perception study

- Might the affective rating achieved for the different voice qualities in the cross-language study be simply an artefact of their differing in loudness?
- Does the inherent loudness of a voice quality contribute to its affective signalling?

The production-based study

This was a descriptive study aimed at illustrating some of the voice quality differences that may be associated with affective renditions in speech. It was hoped this would serve to illustrate (i) the dynamic nature of the kinds of variations which tend to pertain, and (ii) the extent to which these renditions can be characterised by different combinations of voice source parameter settings.

1.8 Structure

The thesis is structured as follows.

In Chapter two, the aim was to provide a theoretical background to the research into voice and emotion. The chapter outlines the main theoretical issues as well as the main empirical finding to date. Main acoustic measures related to voice quality are summarised. Chapter two further discusses the role of voice in communicating affect and outlines the major experimental approaches as well as theoretical models employed in the study of vocal expression of emotion. It further summarises the main findings of cross-cultural studies of the vocal expression of emotions.

Chapter three describes in detail the synthesised stimuli used in the cross-language study as well as the methodology of the auditory experiments. It further describes the statistical procedures used in the data analysis.

Chapter four presents the results of the cross-language study of perception of affect in the synthesised stimuli. A detailed description of results is given for each of the affects in the six subtests followed by the summary of results. Although the main review of the literature is done in Chapter 2, it was considered necessary to include certain references in Chapter 4.

In Chapter five, the role of loudness in affect signalling is investigated in the course of two auditory experiments using synthesised stimuli. The description of the results is preceded by the literature review on loudness.

Chapter six presents a case study of voice source parameter variation in emotion portrayals. The analysis is based on the data obtained by inverse filtering of repetitions of a short all-voiced utterance. The case study illustrates whether the emotions thus portrayed are differentiated in terms of the salient voice parameter settings, and considers some dynamic aspects of the voice source in these portrayals.

Chapter seven presents general conclusions as well as suggestions for future work.

Appendices 1-17 document the details of statistical analyses as well as the data on listeners' agreement in the listening tests described in Chapters 3-5.

2 Chapter two
Voice and emotion

2.1 Voice

'Voice quality as a major vehicle of information about physical, psychological and social characteristics of the speaker has a vital semiotic role to play in spoken interaction' (Laver, 1980, p. 2). An increased interest in the role of voice in human communication and in particular in the role of voice in communicating emotion has given rise to a growing body of research. The sections below provide background for the research presented in this thesis. They outline some common auditory and acoustic measures related to voice quality. The studies of voice and emotion, and in particular cross-cultural studies are further summarised.

2.1.1 Introduction

Voice quality is the focus of study in a broad range of disciplines including acoustics, linguistics and phonetics, medicine, psychology, and speech therapy. A comprehensive discussion of the problems related to defining and measuring voice quality, Kreiman et al. (Kreiman, Vanlancker-Sidtis, & Gerratt, 2003, 2004, 2005) suggest that researchers tend to adopt definitions most suited to their research goals or field of interest. Voice quality can be characterised in a narrow sense as the sound source produced by the airflow through the vibrating vocal folds. Broader definitions consider the contribution of supralaryngeal settings, f_0 and intensity variations, and temporal characteristics of speech as shaping the speaker's tone of voice.

Laver (1980, p. 1) defines voice quality fairly broadly as a quasi-permanent quality, 'the characteristic *auditory* colouring of an individual speaker's voice'. Although Laver's comprehensive description of voice quality includes both the laryngeal (whispery voice, harsh voice) and supralaryngeal settings (pharyngealised voice, labialised voice) as the factors shaping the speaker's tone of voice, intuitively, the auditory percept of a particular voice quality is primarily related to the laryngeal function. In this study, as in most others, I will be referring to voice quality only in terms of the laryngeal source.

As is well established, voice quality is an important and rich source of linguistic, paralinguistic and extralinguistic information. A summary of the many inferences that listeners may make when listening to a voice is given, for example, in Kreiman et al. (Kreiman et al., 2005); see also Gobl and Ní Chasaide (1997). Apart from the verbally

expressed message, voice carries extralinguistic information on the physical characteristics of the speaker. Most characteristics of voice quality result from individual anatomical differences in the size and shape of the vocal apparatus as well as its age, health etc., and are usually beyond the speaker's control. Other more temporary influences will also show such as fatigue, intoxication etc. The social and sociolinguistic environment will also influence a speaker's habitual voice quality settings, and the speaker's language and dialect will leave their imprint (e.g., Esling, 2000; Esling & Wong, 1983).

Some aspects of the linguistic use of voice quality are well studied. The phonologically contrastive use of different phonation types in different languages is discussed in Gordon and Ladefoged (2001), Ladefoged and Maddieson (1996), Ní Chasaide and Gobl (1997). Beyond such contrastive functions, the segmental and prosodic context has major consequences for the voice source characteristics of an utterance. Source parameters vary with different consonants and vowels and there may be large perturbations at the CV transition. Furthermore, there are large dynamic changes in voice source parameters that are part and parcel of the prosodic structure of the utterance. One should note here that these segmental and prosodic effects on the voice source are not heard as changes in voice quality (as noted in Chapter 1 and further below voice quality is a perceptual entity, and the relationship to the acoustic correlates is complex).

In addition to all these strands of information, the voice can reveal or intentionally express a speaker's psychological state (e.g., emotional state, mood). Numerous studies have provided evidence for this ability of the voice to convey paralinguistic information about the emotional state and attitude of the speaker. See, for example, reviews in Johnstone and Scherer (2000), Juslin and Laukka (2003), Juslin and Scherer (2005), Scherer (2003), Laukka (2008), Laukkanen, Alku, Airas and Waaramaa (2008). For instance, in English, breathy voice is claimed to be associated with intimacy and whispery voice is commonly used to signal secrecy and confidentiality (Laver, 1980). Klasmeyer and Sendlmeier (1995, 2000) reported whispery voice in the expression of fear and creaky voice in the expression of sadness. Breathy voice has been found to be associated with the expression of politeness in Japanese spontaneous speech (Ito, 2005). Impressionistic observations of some pathological voice qualities (Behlau & Gasparini, 2008; Izdebski, 2008) suggest certain iconic associations between voice and emotion that exist in Europe and North America. For example, pressed/tense phonation is perceived as angry, hostile, stressed; vocal tremor is associated with nervousness, deception, emotional distress and fear; breathiness can be used purposefully to produce a seductive voice or confidential

voice and may be perceived as such even when produced as a result of phonatory disorder; harsh voice is mainly associated with negative emotions such as fear.

This (paralinguistic) function of voice quality in particular is the focus of this thesis, and a more comprehensive review of literature on the vocal expression/communication of emotion is given in Section 2.3 below.

The main point that needs to be raised here is how difficult it is to establish the correlations of voice quality to affect in speech. The affective/emotional aspects of the voice are part of a complex web of voice source variation which conveys other types of information in parallel. Thus, for example, even if one knows that breathy voice might be associated with sadness, and one extracts a voice source parameter (such as high RD or NAQ) that is generally related to breathiness in some data, we can not necessarily be certain that

(i) this measure is always reliably a measure of breathiness (see Section 2.1.3),

(ii) if it is a reliable measure of breathiness, whether it might indicate something about the intrinsic voice of the speaker, social group, linguistic context, o affective state.

2.1.2 Auditory description

Methods of measuring and quantifying voice quality differ depending on the part of the speech chain the researcher is focused on, whether it is the analysis of vibratory patterns of vocal folds, the study of acoustic signal or the investigation of the perceptual properties of voice. Perceptual characteristics of voice quality are of particular interest if as we and many would argue voice quality is a perceptual phenomenon.

IMPRESSIONISTIC LABELS

Voice quality can be assessed using auditory analysis and defined in terms of specific qualities described by a set of impressionistic labels. Thus a single label is provided for each of a very large number of different voices. Examples of such are given in Laver (1980), e.g., 'heavy voice', 'light voice', 'tinny voice', or 'thick voice'. Scherer (1982b, 1988b, 1992) has suggested the use of scales for rating voice quality and speech characteristics in collecting auditory analysis data from naïve listeners. His scales include such labels as 'thin', 'full', 'tense', 'lax', 'wide', 'clear', 'cultured', 'dark', 'deep', 'thin', 'soothing', 'ringing', 'gloomy', 'flat' etc. The classification of phonation types proposed by Titze (http://www.ncvs.org/ncvs/tutorials/voiceprod/tutorial/quality.html) also makes use of impressionistic labels. Titze's classification is meant primarily for research into

voice disorders and includes such labels as 'flutter' ('bleat'), 'hoarse' ('raspy'), 'honky' ('nasal'), 'resonant', 'strohbass'[1], 'tremerous', 'yawny' and 'twangy'.

As frequently pointed out (Gobl & Ní Chasaide, 2003b; Juslin & Scherer, 2005; Kreiman et al., 2005; Laver, 1980; Ní Chasaide & Gobl, 2005), this approach results in a rather unmanageable descriptive system. Its main drawback is the diverse and somewhat idiosyncratic use of impressionistic terms as the same label may be related to different auditory impressions for different researchers, and different labels may be used by different researchers to describe a single quality.

These rather long lists of labels could be reduced by using statistical procedures, such as factor analysis, or semantic differential scales, yet the application of this approach is still problematic. It requires that a fixed mental representation of the exact quality the label designates exist in the mind of a trained observer as a kind of internal standard. Such internal standards are rarely identical across judges and as a result, a rather low degree of interrater agreement is frequently reported in judgment studies (Kreiman et al., 2005).

LAVER'S PHONETIC DESCRIPTION OF VOICE QUALITY

The most influential taxonomy of voice quality that has been widely adopted in speech research is the one proposed by John Laver in his monograph 'Phonetic Description of Voice Quality' (1980).

Laver (1980) regards both laryngeal and supralaryngeal settings as factors contributing to the speaker's tone of voice in a perceptually important way. Voice quality represents tone of voice characteristics that are quasi-permanent and present practically all the time that a person is talking. This 'background auditory colouring' is independent of the segment-related phonetic variation.

Laver's descriptive model of voice quality is based primarily on auditory impression, and is supported by physiological, acoustic and articulatory data available to date as well as on the analysis performed specifically for the purposes of his monograph. For example, the analysis of acoustic correlates of laryngeal settings was carried out using Sondhi-tube method of analysing glottal volume velocity waveform (Sondhi, 1975) on Laver's imitations of different voices with varying supralaryngeal and laryngeal settings.

An important characteristic of Laver's voice quality taxonomy is that it is based on componential analysis: every particular voice quality is seen as an auditory impression resulting from the combination of independently controllable settings. Such an approach

[1] Extremely low pitched voice; the term is also used in the descriptions of operatic voice, e.g., Henrich (2006)

allows one to describe a considerable number of qualities using relatively few basic voice quality labels.

One important aspect of the Laver system is that it is a system of auditory qualities which people are trained to produce and perceive with accuracy. Like the Cardinal Vowel system, it pins down the potentially infinite range of voice qualities by using clearly defined auditory reference points. This was crucial to the task of trying to define voice qualities acoustically, and reproduce them synthetically.

As this thesis focuses on perceptual study of the vocal correlates of affective states using synthesised stimuli in which glottal source variation was modelled, only the description related to laryngeal settings within Laver's taxonomy will be briefly outlined. Laryngeal settings in Laver (1980) are described in terms of the muscular physiology of the setting, its acoustic characteristics, and phonological and paralinguistic uses of the settings. Laver distinguishes a small number of basic types of phonation, namely, modal, falsetto, creak, breathiness, whisperiness and harshness, and groups them into three categories depending on their ability to exist alone or combine with each other to make up compound phonatory qualities. For example, modal voice and falsetto form one such category because they can exist alone as a simple type or in combination with other types of phonation but not with each other; harshness and breathiness form another group since they cannot occur individually but only in combination with other phonation types. The neutral mode of phonation (modal voice) serves as a baseline against which other modes are contrastively described. It is defined as the phonation type where the vibration of the true vocal folds is periodic, efficient and without audible turbulent noise created at the glottis, 'ordinary voicing' when no specific features attract the listener's attention (Laver, 1980).

Laryngeal settings characteristic of each of the basic phonation types are described in terms of the three parameters of muscular tension interacting with aerodynamic factors of pulmonic airflow: adductive tension, medial compression and longitudinal tension. Adductive tension is defined as the tension of the interarytenoid muscles which brings the arytenoid cartilages together, closing both the cartilaginous and the ligamental glottis. Medial compression is defined as the compressional pressure on the vocal processes of the arytenoid cartilages as a result of contraction of the lateral cricoarytenoid muscles and tension in the lateral parts of the thyroarytenoid muscles. Medial compression closes the ligamental glottis, but not necessarily the cartilaginous glottis whose closing will depend on analytically separate adductive tension. Longitudinal tension is achieved by contraction

of the vocalis and/or cricothyroid muscles (Laver, 1980). Schematic representation of the relationship of these three parameters is given in Figure 2.1.

According to the settings of these three laryngeal parameters, modal voice is produced with moderate adductive tension and moderate medial compression, with moderate longitudinal tension. The fundamental frequency is that of the range used in ordinary conversation. The vibration of the vocal folds is regularly periodic and efficient, without audible frication noise at the glottis.

Whisper is realised with triangular opening of the cartilaginous glottis (inverted Y) about a third of the full length of the glottis. Adductive tension is low, medial compression is moderate to high. The turbulent airflow past the edges of the open cartilaginous glottis results in the characteristic 'whisper' sound generated by friction of the air in and above the larynx. Whisper combined with modal voice gives compound phonation of whispery voice which is characterised by a greater amount of interharmonic noise than in modal voice. In English, a whispery voice is often used to signal secrecy and confidentiality, and this paralinguistic use is likely to occur in other cultures as well (Laver, 1980).

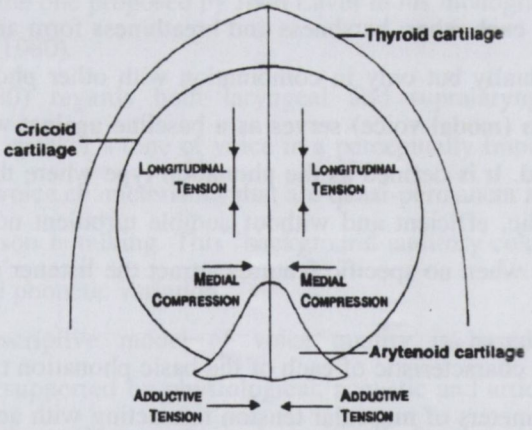


Figure 2.1. *Schematic representation of the action of the three laryngeal parameters. From Gobl (1989), adapted after Laver (1980).*

A characteristic feature of creak (vocal fry, glottal fry; the term 'laryngealisation' is also used) is a very low f_0 (Laver (1980) gives mean f_0 value of 34 Hz). In the production of creak, adductive tension and medial compression are high, but longitudinal tension is low. Neither the length nor the thickness of the vocal folds seems to vary with changes in f_0. Creak combined with modal voice results in the compound phonation type of creaky voice. At the paralinguistic level, creaky voice is used, for example, to signal bored resignation in

English and to express complaint and commiseration in the Mayan language of Tzeltal (Laver, 1980).

Harshness is described as a quality that can only exist in combination with other phonation types. Thus, harsh voice is a modification of modal voice characterised by a combination of extreme adductive tension and extreme medial compression. This excessive approximation of the vocal folds results in irregularity of the glottal waveform and substantial amount of spectral noise, and jitter has been named as the predominant characteristic of harshness. Harsh voice is often used in English as paralinguistic signal of anger.

According to Laver (1980), breathy voice is a modification of modal voice in the production of which the mode of vibration of the vocal folds is inefficient in that the glottis is open along most of its length, the vocal folds never meet on the midline. In breathy voice, the muscular effort is low, adductive tension and medial compression are low, and longitudinal tension is also rather low. Breathy voice is characterised by limited vocal intensity and low f_0. In English, breathy voice is often used for the communication of intimacy.

Laver further points out a close auditory relationship between breathy voice and whispery voice regarding the transition from whisperiness to breathiness as part of an auditory continuum. He suggests the use of the label 'breathy voice' for the range of qualities produced with a low degree of laryngeal effort and where the modal element is markedly dominant over only a slight amount of glottal friction. The term 'whispery voice' can then be used for phonations produced with a greater degree of laryngeal effort, and where the friction component is more prominent than in breathy voice (Laver, 1980). In current speech research, sometimes no distinction is made between breathy and whispery phonation and the results of analysis of the speech signals with aspiration noise are pooled (e.g., Ishi, Ishiguro, & Hagita, 2008).

In his description of voice quality, Laver (1980) further introduces the notion of tension settings related to the overall degree of muscular tension throughout the vocal apparatus. Two types of tension settings are described, tense voice (often described impressionistically as 'bright', 'metallic', 'ringing', 'sharp', 'strident') and lax voice (impressionistically also 'muffled', 'guttural', 'mellow', 'obscure', 'soft', 'thick') which stand for high/low degree of tension respectively throughout the system. The two muscular parameters that are most exploited in the laryngeal contribution to tense and lax voice, along with subglottal pressure, are adductive tension and medial compression.

Laver's system of phonetic description of voice quality has been a basis for one of the most widely used protocols for perceptual evaluation of voice quality - Voice Profile Analysis, e.g., Beck (2005), Laver (2000). The VoQS - Voice Quality Symbols - system used in particular for the phonetic transcription of pathological voices is largely based on it (Ball, Esling, & Dickson, 2000). Another perceptual evaluation protocol often referred to in the literature is the GRBAS scale (Hirano, 1981): Grade (severity of deviation from the norm), Roughness, Breathiness, Asthenia (the auditory impression of weakness in spontaneous phonation) and Strain (vocal tension), e.g., Baken and Orlikoff (2000), Bele (2005), Núñez Batailla et al. (2004).

Laver's taxonomy is highly influential and has been used as standard for the description of voice quality in a large number of studies. K. R. Scherer (1986, p. 145) referred to it as 'a coherent conceptual system for the description of voice quality' which could help significantly advance the study of the role of voice quality in communicating affect. Kreiman et al. (2005) points out that it is the system's limitation that it describes voice quality, essentially a perception phenomenon, in terms of speech production. Furthermore, Laver's description requires the listeners to isolate perceptually important aspects of voice quality on the basis of the auditory analysis, an ability rendered doubtful in the studies by Kreiman and Gerratt (Gerratt & Kreiman, 2000; Kreiman & Gerratt, 2000a, 2000b).

Rather than relying on the existence of fixed internal standards shared by the listener judges, the psychoacoustic methods may be used in the listening tests. For example, in Gerratt and Kreiman (2001) the listeners were asked to adjust the settings of the synthesiser corresponding to the signal-to-noise ratio of the test stimulus in order to match its auditory characteristics to those of the reference voice stimulus. The synthesis settings were then interpreted as parametrically representing the listener's perception of voice quality. The method showed substantial increase in interrater agreement (97% relative to 22% reported in the studies using traditional visual analogue rating scale).

2.1.3 Measuring the acoustic correlates of voice quality

Various acoustic measures have been proposed to provide objective description of voice quality as well as other prosodic parameters relevant to the tone of voice, such as fundamental frequency, duration and intensity. These measures could be divided into two groups: (1) direct methods of analysis of the characteristics of the voice source - the glottal waveform and (2) indirect acoustic measures that analyse the speech waveform. Overviews of common acoustic measures used in the studies of voice quality can be found, for

example, in Gobl (2003), Gobl and Ní Chasaide (2003b), Juslin and Scherer (2005), Keller (2005), Ní Chasaide and Gobl (1997), Scherer, Johnstone and Klasmeyer (2003). Acoustic analysis of voice quality requires simultaneous use of a number of measures as a single measure cannot capture all relevant characteristics of a particular voice quality.

DIRECT VOICE SOURCE ANALYSIS

Voice source data can be obtained by inverse filtering (see description below). Subsequent model matching procedures and parameterisation of the estimated glottal flow is an important part of voice quality analysis. It is an essential tool used in resynthesis-based studies of perceptual characteristics of different phonation types (Gobl & Ní Chasaide, 1999a). It should be noted, however, that direct source analysis is difficult. Although automatic methods of inverse filtering exist that allow one to analyse considerable amount of data, these methods are not accurate. Manual interactive methods allow substantially more accurate estimation of the source signal. However, they require a high level of expertise and can be very time-consuming, and are therefore suitable for the fine-grained analysis of very limited amount of data. Similar problems are related to the parameterisation of the estimated glottal flow obtained by inverse filtering: automatic methods lack accuracy, in particular when parameterisation is based on the time instances in the glottal waveform, such as the instance of opening and closing of the vocal folds. There has been research aimed at using more robust amplitude-based measures (Alku, Bäckström, & Vilkman, 2002; Alku & Vilkman, 1996; Gobl & Ní Chasaide, 2003a; Kane & Gobl, 2009). Manual interactive techniques still remain the method of choice if one desires to get reliable voice source data.

These methods (manual interactive inverse filtering and model matching) were carried out for the utterances analysed/synthesised in Chapter 4 and 5, and for the production study reported in Chapter 6, and will therefore be outlined below. A comprehensive description of inverse filtering and model matching techniques is given in Gobl (2003), Gobl and Ní Chasaide (1999), Ní Chasaide and Gobl (1997), on which the following section is based.

Voice source is generated as a result of modification of the pulmonic airflow through the glottis by the opening and closing action of the vocal folds. In the production of speech sounds, the airflow is further modified by the filtering properties (transfer function) of the supralaryngeal vocal tract. The acoustic source-filter theory of speech production (Fant, 1960/1971) treats the voice source and the vocal tract filter as independent components. The spectrum of the idealised voice source contains all

harmonics and has a constant slope of -12 dB per octave. In speech, the glottal source varies dynamically depending on configuration of the glottis, the degree and type of the laryngeal tension, the respiratory effort and other factors (Ní Chasaide & Gobl, 1997).

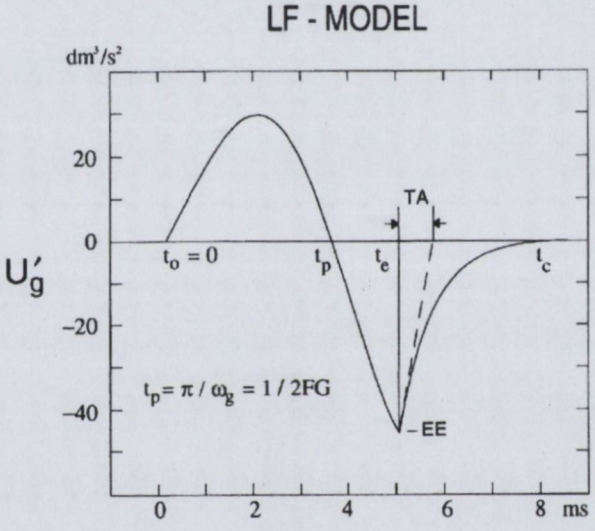
The glottal volume velocity waveform could be obtained by using circumferentially vented pneumatachograph mask for recording oral airflow (Rothenberg, 1973), by reflectionless acoustic tube described by Sondhi (Sondhi, 1975) or by inverse filtering of acoustic speech waveform obtained through high quality recordings using a phase linear microphone.

The process of inverse filtering involves cancelling the filtering characteristics of supralaryngeal cavities by passing the speech signal through a filter whose transfer function is the inverse of that of the vocal tract. The filtering effect of the radiation at the lips is included by using a first order differentiation of the glottal waveform. Inverse filtering can be performed on relatively large amounts of data using automatic and semi-automatic techniques (e.g., Airas, Pulakka, Bäckström, & Alku, 2005), however, as pointed out in Ní Chasaide and Gobl (1997), the most accurate source signal is obtained by using manual interactive methods, for example, Ní Chasaide, Gobl and Monahan (1992) or Gobl and Ní Chasaide (1999). There is a certain trade-off between the amount of data analysed and the accuracy of the measurements: analysis of big databases with automatic inverse filtering will inevitably entail errors whereas a much more accurate analysis of a small data set requires a high level of expertise and could be extremely time consuming.

An estimate of the differentiated glottal flow waveform obtained by inverse filtering is then parameterised by matching it to one of the existing source models outlined, for example, in Cummings and Clements (1995b), Fujisaki and Ljungqvist (1986), Gobl (2003). The voice source characteristics of a given phonation type can be described in terms of voice source parameters obtained from the modelled waveform. A model that has been particularly widely used is the Liljencrants-Fant model - the LF model (Fant et al., 1985). Klatt and Klatt (1990) referred to it as 'superior to other models of the same complexity when the objective is to model natural speech with minimum RMS error' (p. 838).

The LF model consists of two segments (Figure 2.2). The first segment models the differentiated glottal flow from the glottal opening t_o to the time point of main excitation t_e. The segment is represented by an exponentially growing sinusoid the shape of which is defined by three parameters: (1) E_0 - a scale factor required to achieve area balance, (2) $\alpha = -B\pi$, where B is the 'negative bandwidth' of the exponentially growing sinusoid; α determines the increase in amplitude; and (3) $\omega_g = 2\pi F_g$, where $F_g = 1/2t_p$ and t_p is the time

from glottal opening to maximum airflow. The second segment is an exponential function that represents the residual airflow after the main excitation when the vocal folds close (the return phase). It models the differentiated glottal flow from the time point of the main excitation t_e to the glottal closure t_c. The main parameter of the return phase is the time constant TA obtained by the projection on the time axis of the tangent at time point t_e. The LF model therefore determines the shape of the glottal pulse in terms of four parameters, E_0, α, ω_g and TA, along with f_0. The model is further characterised by a requirement of area balance (the positive area of LF pulse should always equal the negative area) (Ní Chasaide & Gobl, 1997)(Ní Chasaide & Gobl, 1997) .



$$U'_g(t) = E_0 \cdot e^{\alpha t} \cdot \sin(\omega_g t) \qquad ; t_o \le t < t_e$$

$$U'_g(t) = \frac{-EE}{\varepsilon \cdot TA} \cdot \left[e^{-\varepsilon(t-t_e)} - e^{-\varepsilon(t_c-t_e)} \right] \qquad ; t_e \le t < t_c$$

Figure 2.2. *The LF model of differentiated glottal flow.*

Important glottal source parameters are defined somewhat differently by different models. Table 2.1 summarises the glottal parameters defined in terms of the LF model, the description is based on Ní Chasaide and Gobl (1997) and Gobl (2003). Figure 2.3 gives examples of a glottal volume velocity flow pulse (upper panel) and corresponding LF model pulse of differentiated glottal flow (lower panel) in which the time points used to calculate the glottal parameters are shown.

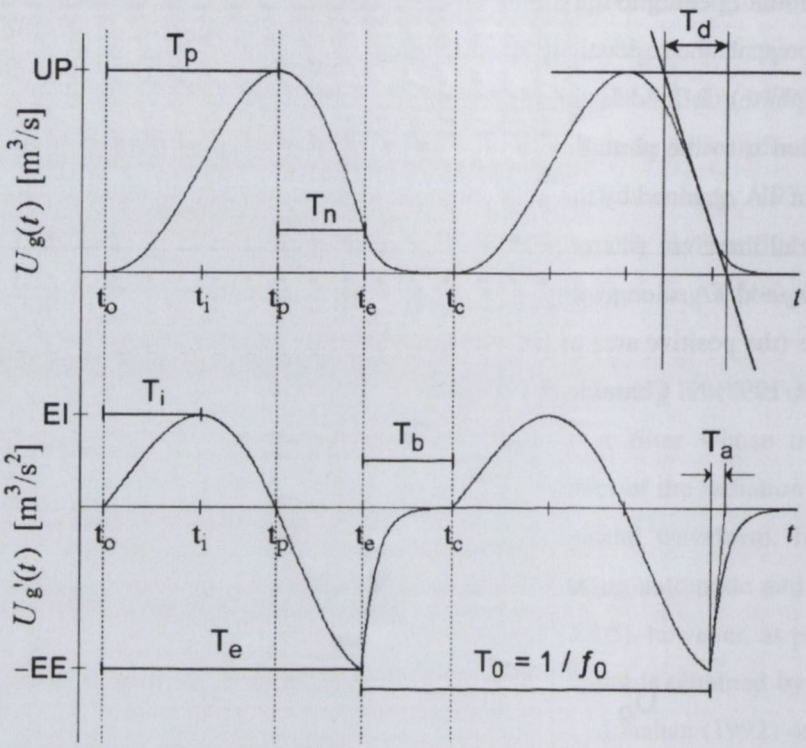


Figure 2.3. *Examples of a glottal volume velocity flow pulse (upper panel) and corresponding LF model pulse of differentiated glottal flow (lower panel) (from Gobl, 2003).*

The source parameters outlined in Table 2.1 tend to covary, for example, the return phase typically varies inversely with EE, so that if the excitation is weaker, RA is higher. There is generally also covariation between RA and RK, so that a long return phase (and low EE) corresponds to a more symmetrical pulse shape (Ní Chasaide & Gobl, 1997).

Phonation types described in Laver (1980) can be characterised in terms of glottal parameter settings. According to Ní Chasaide and Gobl (1997), breathy voice is characterised by high dynamic leakage (high RA values) and a more symmetrical glottal pulse (high RK and OQ values). Tense voice shows opposite characteristics, e.g., very low RA corresponding to a very sharp full closure of the vocal folds. Glottal pulse is more skewed to the right (lowest RK values compared to other phonation types), OQ is small and RG is high.

Table 2.1. *Important glottal parameters defined in terms of the LF model.*

Parameter	Definition	Effects at the acoustic and production level
f_0 (fundamental frequency)	$f_0 = 1/T_0$, where T_0 is the fundamental period	Acoustic correlate of the perceived pitch of the voice
EE (excitation strength)	The negative amplitude value at the point of maximum discontinuity of the differentiated glottal flow.	Main determinant of the overall intensity of the signal. Determined by the speed of closure of the vocal folds. The higher EE, the higher the amplitude of the higher harmonics.
OQ (open quotient)	$OQ = T_e/T_0$, where T_e is the duration of the glottal open phase and T_0 is the duration of the whole glottal period.	Large values of OQ correspond to an increased amplitude of the lowest harmonics in the source spectrum, e.g., in breathy voice.
UP	The peak volume velocity of the glottal pulse.	Large values UP correspond to an increased amplitude of the lowest harmonics in the source spectrum, e.g., in breathy voice
RA (dynamic leakage)	$RA = TA/T_0$, where TA is a measure of the effective duration of the return phase after the main excitation, prior to full or maximum glottal closure, and T_0 is the fundamental period.	It is the residual flow during the return phase from the time of the main excitation to the time of complete or maximum closure. Relates to the sharpness of glottal closure and is therefore an indication of spectral tilt.
FA	$FA = 1/(2\pi TA)$	Captures spectral tilt. A high FA value indicates a source spectrum with relatively strong higher harmonics.
FG (glottal frequency)	$FG = 1/(2T_p)$ where T_p is the duration of the opening phase of the glottal pulse.	FG is the characteristic frequency of the glottal pulse during the open phase. Captures the frequency of the boosting effect of the lowest frequency components of the glottal pulse.
RG	$RG = T_0/(2T_p)$, where T_p is the duration of the opening phase of the glottal pulse $RG = FG/f_0$	RG is a version of the glottal frequency FG normalised to the fundamental frequency. Tends to vary inversely with OQ and UP.
RK	$RK = T_n/T_p$ where T_p is the duration of the opening phase of the glottal pulse and Tn is the duration of the closing phase of the glottal pulse	Captures asymmetry of the glottal pulse. Higher RK values correspond to more symmetrical pulse shapes (e.g., in breathy phonation) while lower RK values suggest a longer closing phase and the glottal pulse skewed to the right.
RD	$RD = 1000 \cdot \frac{UP}{EE} \cdot \frac{f_0}{110} = \frac{1}{0.11} \cdot \frac{T_d}{T_0}$	A global waveshape parameter related to voice quality variation on the tense-lax continuum. RD is expected to be high for breathy voice and low for tense voice.

There has been a growing tendency to introduce measures describing the properties of the glottal waveform with a global parameter, in part owing to the need to analyse relatively big speech databases in terms of voice quality. The most frequently cited ones in the literature are RD (Fant, 1995), AQ and NAQ (Alku et al., 2002; Alku & Vilkman, 1996). These amplitude-based measurements are of particular interest for researchers as they have been suggested as being more robust than the time-domain ones, relatively easy to obtain by inverse filtering methods as well as suitable for analysis of big corpora of speech data (e.g., Campbell & Mokhtari, 2003).

RD is a global waveshape parameter proposed in Fant (1995) and described in the following equation:

$$RD = 1000 \cdot \frac{UP}{EE} \cdot \frac{f_0}{110} = \frac{1}{0.11} \cdot \frac{T_d}{T_0}$$

where T_d is declination time defined by UP (peak volume velocity of the glottal pulse) over EE (the negative amplitude at the time point of maximum discontinuity of the differentiated glottal flow), normalised to the fundamental period. The scaling factor 1000/110 is included so that a numerical value of RD is equivalent to the declination time in ms for f_0 = 110 Hz (Gobl, 2003).

The relationship between RD and other R parameters of the LF model (RA, RG and RK) can be shown to be approximately as follows:

$RD = (1/0.11) \cdot (0.5 + 1.2RK) \cdot (RK/4RG + RA)$ (Fant, Kruckenberg, Liljencrants, & Hertegård, 2000). Fant (1995) suggests that higher RD values correspond to higher RK and lower RG as well as a higher OQ. He further refers to RD as 'the most effective single measure for describing voice qualities' (Fant 1995, p. 152). Applied to the description of different phonation types, RD parameter is expected to be high for breathy voice and low for tense voice, and the range provided in Fant (1995) is $0.4 < RD < 2.7$.

AQ, an amplitude-domain quotient for parameterisation of the glottal source, is defined as the ratio between the amplitude of the AC-flow of the glottal waveform [the level of the maximum flow] and the amplitude of the minimum of the flow derivative [the level of the negative peak amplitude of the flow derivative]: $AQ = A_{AC}/d_{min}$ (Alku & Vilkman, 1996; Alku, Vintturi, & Vilkman, 2001a, 2001b). This essentially corresponds to UP/EE in terms of the LF model parameters. The NAQ parameter introduced in Alku et al. (2002) is AQ normalised to the fundamental period. NAQ is similar to RD, except for the scaling factor. $NAQ = AQ/T_0 = 0.11$ RD (Gobl & Ní Chasaide, 2003a). In general, large NAQ values correspond to smooth glottal flow pulse as in breathy phonation (and low vocal effort). A low NAQ value reflects shortening of the glottal closing phase as in

pressed phonation (and high vocal effort). Normalised amplitude quotient (NAQ) is closely related to the closing quotient (ClQ), another important time-domain parameter used in inverse filtering studies. ClQ is defined as the ratio between the duration of the glottal closing phase and the fundamental period. Airas and Alku (2004), Bäckström, Alku and Vilkman (2002) showed that NAQ is more robust than ClQ.

NAQ has enjoyed relative popularity as a measure suitable for automatic voice quality analysis of big speech databases and its use has been reported in several recent studies, e.g., Airas and Alku (2004), Campbell and Mokhtari (2003), Rossato, Audibert, and Aubergé (2004), Toivanen et al. (2006), Waaramaa, Alku and Laukkanen (2006), Laukkanen et al. (2008).

Gobl and Ní Chasaide (2003a) found that AQ and NAQ measured on samples of emotional speech in Japanese perform differently in different f_0 ranges. AQ is more effective in capturing tense-lax differences across a large f_0 range than the NAQ, which is more suited for a limited f_0 range. They further argued that due to a rather complex correlation between the shape of the glottal pulse and perceived voice quality and f_0 a single glottal parameter is unlikely to capture all aspects of voice quality variation.

INDIRECT MEASURES

Information about voice quality can be also obtained by indirect measures through the acoustic analysis of the speech waveform.

The long-term average spectrum (LTAS) has been reported to reflect voice quality differences, e.g., Banse and Scherer (1996), Bruyninckx, Harmegnies, Llisterri and Poch-Olivé (1994), yet it is not without limitations. As pointed out by Keller (2005), it does not allow differentiation of spectral features relevant to segmental information from those related to the voice source, so the use of LTAS has to be restricted to short steady state portions of vowels. Averaging does not allow capturing of temporal dynamic variation intrinsic to a particular voice quality such as jitter or voice source dynamics characteristics related to segmental variation, such as consonant-vowel transitions. LTAS is an amplitude-sensitive measure in that spectra obtained from high-amplitude signals differ considerably from those obtained from low-amplitude signals. Finally, averaging methods have shown only very weak differentiation among phonation types (Keller, 2005). LTAS is used for other spectral measures such as alpha ratio which describes the spectral energy distribution and thus the voice quality along the tense-lax continuum and is calculated by subtracting the SPL in the range 50 Hz-1 kHz from SPL in the range 1-5 kHz (Toivanen et al., 2006).

One of the major acoustic parameters that have been reported to reliably differentiate phonation types is spectral tilt, i.e. the degree to which the amplitude of

spectral components decreases with increased frequency[2]. Some common measures of spectral tilt are summarised in Table 2.2. Spectral tilt can be quantified, for example, by comparing the amplitude of the fundamental (H1) to that of higher frequency harmonics. A very dominant H1 has been reported as characteristic of breathy voice, and subtraction of the amplitude of the H1 from the amplitude of higher harmonics gives the highest values for creaky voice and the lowest values for breathy voice. H1-H2 measure has been reported to correlate with the open quotient (higher for breathy voice) and with the abruptness of vocal fold closure (potentially associated with creaky voice) (Erickson, 2005; Gordon & Ladefoged, 2001; Keating & Esposito, 2007; Ladefoged, 2003). However, these types of measurement are sensitive to the vocal tract filter characteristics as well as to shifts in the f_0 to F1 ratio and may not represent valid measures of spectral slope (Ní Chasaide & Gobl, 1997).

A set of acoustic parameters of the voice source related to spectral tilt that are corrected for the boosting effect of the vocal tract transfer function (as shown by asterisks) have been proposed in Hanson (1997), Hanson and Chuang (1999), see Table 2.2. These corrected measures of spectral slope have been used, for example, in Iseli, Shue and Alwan (2007), Iseli et al. (2006) to study the effect of age and sex and vocal tract configuration on the glottal excitation signal in speech. Kreiman, Gerratt and Antoñanzas-Barroso (2007) referred to Hanson's corrected values as 'a sort of virtual inverse filtering that does not require the use of specialized recording equipment' (p. 599).

Fant and colleagues (Fant, 1997; Fant, Kruckenberg, Liljencrants et al., 2000) used the difference (in dB) between the high frequency pre-emphasised measure SPLH and the sound pressure level SPL as a measure of spectral tilt. They argue that at constant articulation, variations in the SPLH-SPL correspond to variations in the high frequency contents of the source, which in turn is related to the concept of spectral tilt.

Other measures of spectral slope include the parabolic spectral parameter PSP (Alku, Strik, & Vilkman, 1997) measured directly on source pulses obtained by inverse filtering. PSP value is a number that expresses, using a parabolic function, the spectral decay of the analysed voice source with respect to its maximal theoretical decay.

[2] Heldner (2003) commented on the diversity in the literature as regards the terms denoting the slope of the spectrum (e.g., spectral tilt, spectral balance or spectral slope) as well as the methodology in measuring it. He suggested that two classes of measures be distinguished, with the term 'spectral tilt' explicitly representing the slope of the spectrum and 'spectral emphasis' used for measuring the relative contribution of the high-frequency parts of the spectrum to the overall intensity. These measures are not identical; e.g., an increase in spectral emphasis results in a decrease in spectral tilt.

Table 2.2. *Some acoustic measures of spectral tilt relevant to voice quality analysis.*

Parameter	Definition	Effects at the acoustic and production level
H1-H2; H1-F1; H1-F2 e.g., Erickson (2005) Gordon and Ladefoged (2001) Keating and Esposito (2007) Ladefoged (2003)	Amplitude of the fundamental (H1) relative to that of the higher harmonics, e.g., the second harmonic (H2), the harmonic closest to the first formant (F1), or the harmonic closest to the second formant (F2)	A very dominant H1 is a characteristic of breathy voice; the difference between the amplitude of the H1 and the amplitude of higher harmonics is the highest for creaky voice and the lowest for breathy voice; H1-H2 measure correlates with the open quotient (higher for breathy voice) and with the abruptness of vocal fold closure (potentially associated with creaky voice)
H1*–H2* Hanson (1997) Hanson and Chuang (1999)	Amplitude of the first harmonic (H1) relative to that of the second (H2), corrected values as shown by asterisks	An indication of the open quotient Spectral tilt
H1*–A1 Hanson (1997) Hanson and Chuang (1999)	Amplitude of the first harmonic (H1) relative to that of the first formant peak in the spectrum (A1), H1 corrected for the effect of the first formant as shown by asterisk	An indication of the presence of a posterior glottal chink characteristic of whispery phonation Spectral tilt
H1*–A3* Hanson (1997) Hanson and Chuang (1999)	Amplitude of the first harmonic (H1) relative to that of the third-formant spectral peak (A3), H1 is corrected for the effect of F1, and A3 is corrected for the F1 and F2	Indication of the degree of abruptness of the glottal closure Spectral tilt
SPLH-SPL Fant (1997); Fant, Kruckenberg, Liljencrants and Hertegård (2000)	The difference (in dB) between the high frequency pre-emphasised measure SPLH and the sound pressure level SPL	Correlated with variations in the high frequency contents of the source Measure of spectral tilt
The parabolic spectral parameter (PSP) Alku, Strik and Vilkman (1997)	A mesure of the spectral decay of the analysed voice source with respect to its maximal theoretical decay	Spectral tilt
Harmonic Richness Factor (HRF) Childers and Lee (1991)	A measure of amplitude relationships between the fundamental and the higher harmonics	Spectral tilt

Childers and Lee (1991) defined a parameter to measure amplitude relationships between the fundamental and the higher harmonics, Harmonic Richness Factor (HRF), that they used in parameterisation of the glottal flow in the frequency domain in the study of acoustic correlates of vocal folds vibratory patterns in modal, creaky, tense and breathy voices. Another possible measure of spectral slope is the deviation of the empirical source slope from an 'ideal' slope of -12 dB per octave (Kreiman, Gerratt, & Antoñanzas-Barroso, 2007; Ní Chasaide & Gobl, 1997).

Kreiman et al. (2007) argued that it remains unclear how different measures of the source spectral slope relate to each other. In an experimental study, they examined a number of existing measures of the glottal source spectrum, to determine the measures best capturing information about the shape of glottal pulses and the source spectra. The measures were applied to three different types of signals: the harmonic part of the glottal source (estimated by inverse filtering), the complete glottal source (including the inharmonic component), and the complete oral speech signal. Principal component analysis reduced the original 78 measured parameters to four independent factors: Hl-H2, overall spectral slope, high-frequency noise excitation, and H2-H4. The authors report high levels of duplication among existing measures of the glottal source spectrum. Compared to other measures, H1-H2 has been found to be particularly robust as a measure of the shape of both the spectrum and the glottal pulse. H1-H2, the spectral slope in the mid-frequencies (about 1-3 kHz), and high-frequency excitation (harmonic and inharmonic) are reported to account for more than 25 per cent of variance in spectral shapes, suggesting that these features may be important determinants of voice quality (Kreiman, Gerratt, & Antoñanzas-Barroso, 2007).

Frequently used acoustic measures intended to reflect variation in perceived voice quality are f_0 and amplitude perturbation measures jitter and shimmer as well as the measures of the spectral noise Harmonic-to-Noise Ratio (HNR) or Noise-to-Harmonic Ratio (NHR). It has been frequently pointed out, however, that these measures, although relatively easy to obtain, are not sufficient for an exhaustive description of voice quality (e.g., Baken & Orlikoff, 2000). Glottal-to-Noise Excitation Ratio (GNE Ratio) based on the correlation coefficient for Hilbert envelopes of different frequency bands was introduced in Michaelis, Gramss and Strube (1997). GNE Ratio has been shown to be a robust measure of the relative noise level in the speech signal and almost independent of jitter and shimmer. Combining jitter, shimmer, and GNE measurements has been suggested in Michaelis et al. (1997) as a way to obtain a more reliable objective description of voice quality.

To conclude, voice quality represents a complex phenomenon in terms of both conceptualisation and measuring and there appear to be a number of definitions and analytical approaches that need to be treated as complementary. One parameter is unlikely to capture all relevant characteristics of voice quality (such as, for example, its affective content) and so a number of measures need to be included if the description aspires to be comprehensive. Voice quality in itself needs to be treated first and foremost as a perceptual entity, and the study of its physical characteristics should be based on the auditory analysis. Just because we measured a parameter that is generally associated with breathiness does not mean that a particular chunk of speech is heard as having breathy voice quality.

Ní Chasaide and Gobl (2004b) suggest that a holistic approach to the study of the voice source need to be adopted in which f_0 and voice quality interaction, along with the temporal characteristics of speech, is considered. Very few publications reflect this kind of integrated approach, e.g., Epstein (2002, 2003), Pierrehumbert (1989), Swerts and Veldhuis (2001), and obviously more work is required here. Note that in this thesis, voice qualities are treated as holistic entities that are tested in terms of their ability to impart affective colouring.

2.1.4 Vocal expression of emotion

The vocal expressions are fundamental in social interaction. Johnstone and Scherer (2000), Juslin and Scherer (2005) describe vocal expressions as an evolved physiological mechanism that with time was shaped by the social environment into formal signals. Spontaneous vocal expression of affect is determined mainly by the physiological changes that accompany emotional episodes, which in turn influence the mechanism of voice production. For example, in rage, increased tension in the laryngeal muscles and raised subglottal pressure will influence the vibratory pattern of the vocal folds and thus voice quality in a certain way. In real life situations of communication spontaneous expression are constrained by the socially acceptable norms or 'display rules'. In acted speech, in which emotion-specific vocal expression are largely controlled by the performer, these subtle physiologically-induced changes are also expected to be absent. In Scherer's terminology, affect-related change in voice are influenced by both 'push effects' (emotion-related physiological processes, e.g., rate of respiration and facial muscle tension) and 'pull effects' (external factors related to socially accepted norms or listener's expectations, speaker's self-presentation etc.) (Johnstone & Scherer, 2000; Juslin & Scherer, 2005).

The ability of voice to carry information on the speaker's emotional state, mood and attitude has been the focus of a number of studies, as well as being central to this thesis.

The interest in the subject has grown noticeably over the past 30 years, and the amount of empirical data collected is considerable. A number of reviews and meta-analytical studies summarised the existing theoretical models and empirical evidence on the effects of emotions on voice, for example, Bachorowski and Owren (2003), Cowie and Cornelius (2003), Elfenbein and Ambady (2002b), Erickson (2005), Frick (1985), Gobl and Ní Chasaide (2003b), Juslin and Laukka (2001, 2003), Juslin and Scherer (2005), Murray and Arnott (1993), Laukka (2008), Pittam and Scherer (1993), Scherer (1986, 2003), Scherer et al. (2003).

The study of the role of voice quality in communicating affect is a multidisciplinary field, which results in the diversity of methodologies and theoretical issues. A theoretical model that has been suggested for the study of voice and emotion is the Brunswik lens model (Figure 2.4), e.g., Scherer (1978, 1982b, 1986, 2003), Scherer et al. (2003). The Brunswik lens model was systematically implemented in some recent studies of vocal expression of affect, e.g., Bänziger (2004), Bänziger and Scherer (2003), Juslin and Laukka (2001), Grandjean, Bänziger and Scherer (2006) and is outlined below.

The Brunswik lens model represents the non-verbal communication process as follows. The emotional arousal of the speaker is assumed to be accompanied by physiological changes that affect mechanisms of speech production resulting in emotion-specific patterns of acoustic cues. The internal emotional state of the speaker is thus encoded in certain objectively measurable speech characteristics ('distal cues'). These acoustic parameters are transmitted to and perceived by the listeners. Some of the perceived cues (referred to as 'proximal cues' in the model) are used by the listeners to infer the emotional state of the speaker. Objectively measurable acoustic properties of the emotionally coloured speech (distal cues) are not necessarily equivalent to the proximal cues as perceived by the listener. Distal cues may be modified or distorted by the transmission channel and the properties of the hearing system. The model makes a clear distinction between the stages of the process of vocal expression and communication of emotion (Bänziger & Scherer, 2003; Juslin & Scherer, 2005; Scherer, 2003).

An important feature of the Brunswik lens model is that it views the emotion communication process as cue integration: listeners combine several different voice cues to achieve correct emotion inference. Furthermore, a larger number of voice cues used results in a more reliable communication of emotion. On the other hand, redundancy of vocal cues implies that different cues may substitute for one another. Individual voice cues may not always be reliable indicators of the underlying emotion as, for example, the same cue may

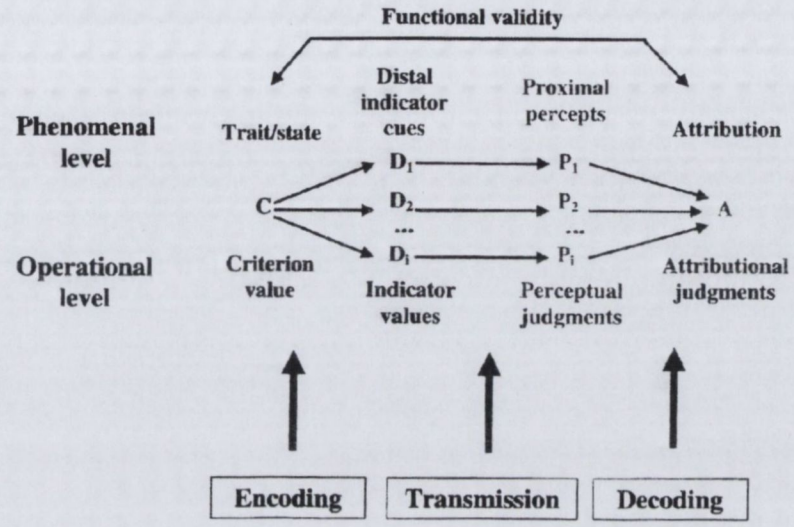


Figure 2.4. *The Brunswik lens model in the description of the vocal communication of emotion (from Scherer et al., 2003).*

be associated with more than one emotion. Therefore, many cues need to be considered in order to thoroughly characterise encoding/expression of an emotion (Juslin & Laukka, 2003; Juslin & Scherer, 2005).

The Brunswik lens model interprets vocal signalling of emotion in a somewhat similar way to the speech chain model (Denes & Pinson, 1993) which incorporates articulatory, acoustic and auditory stages of the speech communication process.

2.2 Emotion

This thesis studies the role of voice quality and communication of affect in the course of listening tests in which listeners rate synthesised stimuli for the presence and strength of affective colouring on a number of bipolar scales. Defining emotion and affect for this study was therefore considered as important as defining the voice: we define what (emotion/affect) is measured in or associated with voice. The selection of the anchor labels of the rating scales has certain implications for the interpretation of results of such associations. Working definitions of emotion and affect as well as the most commonly used models are outlined in the sections below.

2.2.1 Terminology and definitions

As frequently pointed out, e.g., Cowie and Cornelius (2003), Fehr and Russell (1984), Juslin and Scherer (2005), Plutchik (1980), Russell (2003), emotion is a phenomenon exceptionally difficult both to define and to study, and there is still no agreement in the research community with this regard. Scherer (2004) attributes this difficulty to the 'extraordinary fluidity and changeability of the emotion process, which is often hard to segment into discrete chunks and even harder to describe with discrete language labels (which denote steady states and ideal categories)' (p.669). He further suggests the use of the term 'emotion episodes' as more appropriate compared to 'emotion states' as it emphasises the dynamic nature of emotion (Scherer, 2000b). Plutchik (1980, p. 22) points out that 'the history of psychology is characterised by so many different meanings of the term emotion that some psychologists have suggested the word be eliminated from psychological writing'. Russell (2003, p. 145) makes similar observation as regards the absence of formal criteria for what is and what is not an emotion: 'I am often left puzzled by just what a writer means when the word *emotion* appears in a statement or printed in a box in a flow chart. Also puzzling is the array of vastly different theories about the nature of emotion that have arisen. Indeed, few writers have failed to compare emotion as described by psychologists with the elephant as described by blind men in an old fable. Are emotions to be conceptualized as brain modes, actions or action tendencies, reflexes, instincts, attitudes, cognitive structures, motives, sensations, or feelings? Are they biologically fixed modules (and hence reducible to biology) or socially constructed roles

(and reducible to sociology)? Discrete categories or bipolar dimensions? Cognitive, precognitive or postcognitive?'

Further discussion of the definitions of emotion-related terms can be found, for example, in Banse and Scherer (1996) , Cowie and Cornelius (2003), Fehr and Russell (1984), Guerrero, Andersen and Trost (1998), Kleinginna and Kleinginna (1981), Murray and Arnott (1993), Oatley, Keltner and Jenkins (2006), Plutchik and Kellerman (1980), Russell and Barrett (1999), Scherer (2000c), Douglas-Cowie, Cowie, Martin, Devillers and Cox (2006), Scherer and Brosch (2009). The definitions most frequently occurring in the literature on vocal expression of emotion will be outlined here, mainly following Juslin and Scherer (2005) and Scherer (2003). These definitions are also adopted in this thesis.

Affect is a general term covering a variety of phenomena such as emotion, stress, mood, interpersonal stance (attitude), affective personality traits all of which represent a certain departure relative to the neutral state. Affective states broadly fall into three classes depending on intensity, duration and rapidity of change, 'degree of synchronisation' of the systems of the organism in reaction of the eliciting event as well as on the type of elicitor itself: (1) emotions and stress, (2) moods and interpersonal stances, (3) attitudes and affect dispositions (e.g., personality). The design-feature delimitation approach suggested by Scherer (2000c, 2003) allows clear differentiation among these affect-related phenomena. It is reproduced in Table 2.3 below.

Emotions are defined as short intense episodes of coordinated changes in several components (the 'emotion reaction triad': physiological arousal, motor expression and subjective feeling; and possibly also action tendencies and cognitive processes) in response to events that are of major importance to the individual. The expression of true emotions is spontaneous and does not require an 'audience', that is it can exist independently of communication.

Moods and interpersonal stances are rarely provoked by anything specific and are not based on appraisal. Moods may occur for reasons often unknown to the individual, e.g., fatigue, hormonal balance, and weather. Interpersonal stances are defined as long lasting and slow to change intentional or unconscious affective styles in which a person interacts with other people. Attitudes are low intensity long term affective evaluations of objects or persons (Juslin & Scherer, 2005; Scherer, 2003). Expression of attitudes therefore presupposes interaction: you can be happy on your own, but not condescending or arrogant on your own. The list of affective states described by Scherer's design feature approach has been further expanded by HUMAINE project (Douglas-Cowie et al., 2006), mainly for

Table 2.3. *Design feature delimitation of different affective states. Symbols indicate the degree to which the features are present, with 0 indicating the lowest (absence) and +++ indicating the highest; arrows indicate hypothetical ranges. From Scherer (2000c).*

Brief definition of affective states	Intensity	Duration	Synchro-nisation	Event focus	Appraisal elicitation	Rapidity of change	Behavioural impact
Emotion: relatively brief episode of synchronised responses by all or most organismic subsystems to the evaluation of an external or internal event as being of major significance (e.g., anger, sadness, joy, fear, shame, pride, elation, desperation)	++→+++	+	+++	+++	+++	+++	+++
Mood: diffuse affect state, most pronounced as change in subjective feeling, of low intensity but relatively long duration, often without apparent cause (e.g., cheerful, gloomy, irritable, listless, depressed, buoyant)	+→++	++	+	+	+	++	+
Interpersonal stances: affective state taken toward another person in a specific interaction, colouring the interpersonal exchange in that situation (e.g., distant, cold, warm, supportive, contemptuous)	+→++	+→++	+	++	+	+++	++
Attitudes: relatively enduring, affectively coloured beliefs, preferences, and predispositions toward objects or persons (e.g., liking, loving, hating, valuing, desiring)	0→++	++→+++	0	0	+	0→+	+
Personality traits: emotionally laden, stable personality dispositions and behaviour tendencies, typical for a person (e.g., nervous, anxious, reckless, morose, hostile, envious, jealous)	0→+	+++	0	0	0	0	+

the purposes of naturalistic data labelling, to include differentiation between established and (full-blown or suppressed) emergent emotion, stance towards object/situation and stance towards person, interpersonal bonds etc.

Some emotions are further described as biologically-based, hardwired, and innate; they are deemed to be expressed and recognised universally. Other 'complex' affective states involve a significant cognitive component; they are described as having evolved in the course of social interaction, and are more likely to manifest culture-specific characteristics. They require skills of 'mind reading' or 'social intelligence' (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001) that involves attribution of a cognitive mental state to oneself or another person. Studies by Baron-Cohen and colleagues (Baron-Cohen, 2001a, 2001b; Baron-Cohen et al., 2001) suggest that emotion recognition of the former (happiness, sadness, anger, fear, surprise, disgust) is relatively unaffected in people with autism, whereas complex 'mental states' (such as thoughtful, admiring, flirting, bored, interested and arrogant) is significantly impaired.

A problem frequently encountered by emotion researchers concerns the description of emotion and affect in terms of affective labels or affect-denoting words ('emotion lexicon'), e.g., Davitz (1969), Fehr and Russell (1984), Greasly, Sherrard and Waterman (2000), Russell and Mehrabian (1977), Shaver, Schwartz, Kirson and O'Connor (1987). There have been attempts to provide exhaustive list of emotion-denoting words. According to Cowie and Cornelius (2003), there is broad agreement on the scale of the emotion lexicon: different authors list from 100 to 500 emotion-related words, the most common number being about 250, and a subset of prototypical terms in common use amounts to 60-70 words. A list of 89 emotion-denoting words frequently used in the studies of vocal expression is given, for example, in Juslin and Scherer (2005, p.94)[3]. A taxonomic system of Mind Reading[4] (Baron-Cohen, 2007) includes 412 emotions and mental states, grouped into 24 emotion groups. A core list of emotion words compiled for the HUMAINE project (Douglas-Cowie et al., 2006) consists of 48 words.

To reduce these rather long lists, statistical procedures such as cluster analysis or the intuition of lay judges have been used. For example, Cowie et al. (1999) used the opinion of naïve judges to identify a small vocabulary of words sufficient for the description of most emotional states in everyday life (to which they refer as a Basic English Emotion Vocabulary). The subjects were asked to select, from the initial

[3] According to their data, the most frequently occurring emotion terms are anger (used in 85 studies), fear (65 studies), sadness (65), happiness (44), joy (31), disgust (26), surprise (24), boredom (17), and contempt (15).
[4] An interactive guide to emotions and mental state designed to help people with autism to recognise emotions from face and voice.

vocabulary of 45 terms compiled from published lists of emotional terms, sixteen words that they considered as sufficient for the description of emotion in everyday life situations. The words most frequently selected by the majority of subjects were happy, angry, sad, interested, pleased, relaxed, and worried. A similar approach was used in Fehr and Russell (1984) who asked their subjects to list words that fit under the general category of 'emotion'. Their results showed that the words most frequently appearing in the lists are happiness, sadness, love, fear, hate, and joy. The list of emotions in Mind Reading (Baron-Cohen, 2007) was compiled on a different principle: it includes 'any word that describes how someone could feel or how their face could look or how they voice could sound providing the word has an emotional dimension'. Every emotion falls into one of 24 families/groups of emotions. For example, the HAPPY group includes *comfortable, glad, happy, joking, lucky, merry, safe,* and *sociable.*

Most of the affective labels used in the listening tests described in Chapters 4 and 5 and discussed further in Chapter 3 (Methodology) include the words most frequently occurring in the literature: *happy, sad, bored, interested, relaxed, stressed, scared*, and *indignant.* Terms that do not appear often were selected for various reasons – either to ensure the bipolarity of the scale, e.g., *fearless* vs. *scared, apologetic* vs. *indignant* or to represent affects often associated with a particular voice, e.g., *intimate* (often associated with breathiness) vs. *formal.*

2.2.2 Theories of emotion

Plutchik (1980) distinguishes four major theoretical approaches in the literature on the nature of the emotion: (1) the evolutionary tradition of Charles Darwin, that views emotions as evolved types of adaptive behaviour, some innate and some learned, with important survival functions; (2) the psychophysiological tradition of William James/Karl Lange, that claims that emotion is the feeling of bodily changes that directly follow a perception of an eliciting event ('we feel sorry because we cry'); (3) the neurological tradition, mainly focused on the relation between emotional feelings or behaviours and brain structures and activities, and (4) the dynamic tradition and psychoanalysis that emphasises the idea that some aspects of emotions are unconscious and cannot be examined by introspection.

Prinz (2004a) divides emotion theories according to the kind of emotional component(s) (such as bodily changes, action tendencies, modulation of mental processes such as attention, and conscious feelings) with which a particular theory identifies emotion. Thus, 'somatic' and 'feeling theories' (e.g., of James/Lange tradition) identify emotion

with conscious experience and changes in the body and face, behavioural theories identify it with action tendencies, 'processing mode theories' identity emotions with systematic changes in functioning of memory, attention, reasoning, and for 'pure cognitive theories' emotions are identical to thoughts.

In a review of the most influential theoretical perspectives on emotion in psychology, Cornelius (2000) names, apart from the Darwinian and Jamesian traditions, two other prominent approaches: cognitive and social constructionism/constructivism. The cognitive approach posits that emotions arise from an individual's cognitive appraisal of a situation and its implication for personal well-being; different appraisals of eliciting events lead to the experience of distinct emotions. Social constructivism regards emotions as 'culturally scripted behaviours' governed by rules that define how, when and by whom particular emotions are to be experienced and expressed (Cornelius, 2000).

The complementary nature and increasing integration of these approaches is pointed out, for example, in Guerrero et al. (1998) and Cornelius (2000). Each theoretical perspective provides unique information about the structure of emotion and each can be applied to speech communication research.

Reviews of different theories of emotion could be found, for example, in Andersen and Guerrero (1998), Burleson and Goldsmith (1998), Cornelius (2000), Guerrero et al. (1998), Oatley et al. (2006), Plutchik (1980), Prinz (2004a), Scherer (1999, 2000b, 2000c). In emotional speech research, three models have been most frequently applied, namely: (1) discrete or basic emotions, (2) dimensional theories; and (3) appraisal theories, and they will be briefly outlined below.

DISCRETE EMOTIONS (BASIC EMOTIONS)

Basic emotions are defined as a limited number of 'major adaptive emotional strategies' developed during the course of evolution. Each of the basic emotions is characterised by its own 'specific eliciting conditions and its own specific physiological, expressive and behavioural reaction patterns' (Scherer, 2000c, p. 147). The discrete emotion models are based on Darwin's 'The Expression of the Emotions in Man and Animals' (Darwin, 1872/1999) and have been developed in the works of S. S. Tomkins, P. Ekman and C. E. Izard. Tomkins defined basic emotions as phylogenetically stable neuromotor programs and suggested that there are eight of them: interest, surprise, joy, anguish, fear, shame, disgust, and rage (Tomkins, 1995). The work of Ekman and Izard (Ekman, 1992, 1999, 1982; Ekman et al., 1987; Izard, 1990, 1994) showed that there are universal prototypical patterns of facial expressions for a number of basic emotions such as joy, surprise, anger, disgust, contempt, shame and fear. These emotions are expressed

similarly across different cultures and are universally recognised[5]. Basic emotions are hypothesised to combine to form 'emotional blends'; e.g., rage may be conceived of as a blend of anger and fear. Discrete basic emotions are highly unlikely to occur in their pure form. According to Schröder (2004a), in databases of spontaneous emotional speech these basic emotions are rarely found.

The existence of emotion-denoting words with a very high frequency of usage, such as anger, fear, sadness and joy, support the idea of discrete emotions (Scherer, 2000c; Scherer et al., 2003). From a linguistic perspective, basic emotions could be defined in terms of 'semantic primitives' (Johnson-Laird & Oatley, 1989). 'Semantic primitives' such as happiness, sadness, anger, fear and disgust are assumed to be impossible to define using other words and therefore represent emotions at the most basic level. Second-level basic emotions are derivatives of the five semantic primitives, e.g., love is a special case of happiness and hate is a special case of anger (Guerrero et al., 1998).

Although various authors give different number of basic emotions (for example, Scherer (2000c) gives a range of seven to fourteen), most lists of basic emotions include anger, happiness, sadness and fear. The 'Big Six' commonly studied in emotion and voice research includes surprise and disgust. Ekman (1999) extended the list to include amusement, contempt, contentment, embarrassment, excitement, guilt, pride in achievement, relief, satisfaction, sensory pleasure, shame.

The existence of discrete emotions implies that there exist distinct and predictable emotion specific patterns of combinations of vocal parameters (Scherer, 2000c; Scherer et al., 2003). The results of experimental studies reported in Laukka (2004) explicitly suggest that a discrete emotion framework provides the best account of vocal expression of emotion. However, this approach has been much criticised (L. F. Barrett, 2006; Ortony & Turner, 1990; Planalp, 1998). For example, Ortony and Turner (1990) argued that basic emotions out of which all other emotions are built (or through which all other emotions can be explained) 'represent an article of faith rather than an empirically or theoretically defensible basis for the conduct of emotion research' (p. 315). Cowie and Cornelius (2003) suggest that lists of basic emotions need to be treated with caution as each entry actually stands for a family of emotions, often with quite dissimilar members.

[5] Studies by Baron-Cohen et al. (Baron-Cohen et al., 1996; Baron-Cohen, Wheelwright, & Jolliffe, 1997) showed that accurate cross-cultural recognition of emotion from face extends beyond basic emotions to more complex 'mental states' (e.g., scheme, revenge, guilt, regret, distrust).

DIMENSIONAL MODELS OF EMOTIONS

Dimensional models categorise emotions according to a number of dimensions. Unidimensional models described in Scherer (2000c) use one such dimension (usually, activation), multidimensional models usually include three, e.g., valence (i.e. positive or negative affect), activation (i.e. passive - aroused) and intensity (i.e. strong - weak). The three dimensions proposed in Schlosberg (1954) are pleasantness/unpleasantness, attention/rejection, intensity or level of activation. In Russell (1991), the concept of emotion can be defined using three bipolar dimensions: (1) evaluation, valence, pleasure or positivity; (2) activity, arousal or activation; (3) potency, power, control or dominance.

Dimensional models using valence and activation have been particularly popular in emotional speech research, e.g., Bachorowski (1999), Bachorowski and Owren (1995; 2003), Grimm, Kroschel, Mower and Narayanan (2007), Schröder (2004a, 2004b), Schröder, Cowie, Douglas-Cowie, Westerdijk and Gielen (2001), Cowie et al. (2000). Murray, Arnott and Rohwer (1996) point out that basic emotion and dimensional theories are not mutually exclusive, as discrete emotions can be considered as points within a dimensional model.

According to Guerrero et al. (1998), the most widely used dimensional models of emotions are Russell's circumplex model, a three dimensional model, and Plutchik's multidimensional model.

In Russell's two-dimensional circumplex model (Russell, 1980), emotions are plotted on a two-dimensional circular structure (circumplex) based on their level of activation (i.e. passive vs. active) and their valence (i.e. positive or negative) (see Figure 2.5). For example, 'bored' is characterised by low activation and negative affect, whereas 'delighted' is characterised by moderately high activation and positive affect. Emotions that are across the circle from one another (i.e. alarmed and calm) correlate inversely. Russell (2003) argues that valence and arousal are the 'core processes' of affect and constitute the primitive basis of emotional experience. Larsen and Diener (1992) point out that, although clearly structured, the model may be too simplistic to capture the complexity of emotion (e.g., valence and activation are not enough to describe and differentiate anger and fear). Moreover, some of the emotion terms within the model do not fit the definition of emotion (e.g., sleepy and tired).

It should be noted that the model has been extensively used in the studies of emotion in speech and that it is largely supported by empirical evidence: acoustic correlates of affect in speech relate mainly to dimensions of activation and valence.

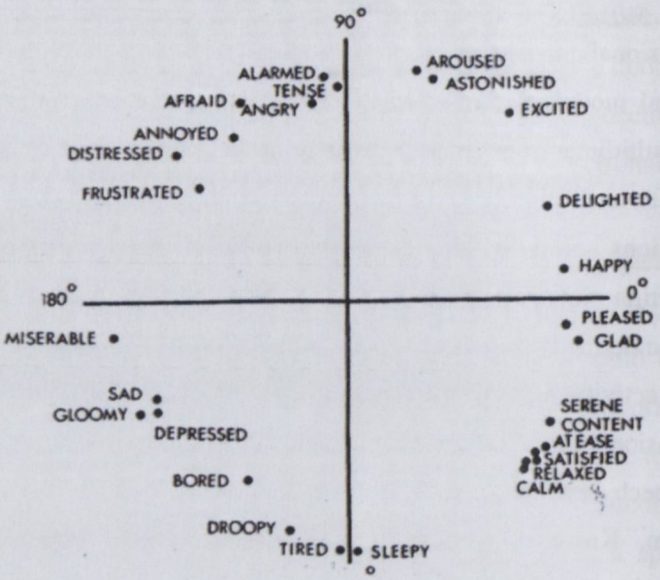


Figure 2.5. *Russell's circumplex model of affect (from Russell, 1980).*

The two dimensions of valence and activation can be extended by introducing a third dimension. A three dimensional conical model was suggested, for example, in Daly, Lancee and Polivy (1983) who added emotional intensity dimension to the circular organisation of dimensions of valence and activation. Affective labels are arranged on the surface of the cone according to their intensity, and the top of the cone represents the neutral state. Thus, 'happy' will be placed closer to the base of the cone compared to 'pleasant'. In their model, intensity is not identical to activation. Intensity is concerned with how strong or weak the emotion is and shows a change from a neutral state in a non-directional manner. Activation is related to deviation from one's normal physiological state; it decreases or increases (the directionality is up or down). Emotions could be characterised by low activation and high intensity. For example, depression is characterised by negative valence, low activation and high intensity (Daly et al., 1983; Guerrero et al., 1998).

Plutchik's multidimensional model (Plutchik, 1980) describes basic emotions from which all other emotions can be derived by arranging a set of eight basic emotions into a three dimensional emotion cone (or half-orange), with the emotion-denoting words corresponding to emotion at maximum intensity (e.g., rage, grief, admiration etc.) are at the top (Figure 2.6). The vertical dimension represents intensity, or level of arousal, and ranges from a maximum state of excitement to a state of deep sleep at the bottom. The shape of the model implies that the emotions become less distinguishable at lower intensities.

Within the model, primary emotions may be mixed to produce complex emotions: any adjacent pair of primary emotions could be combined to form intermediate mixed emotions - dyads and triads (combinations of two or three primary emotions), with once or twice removed primaries forming secondary or tertiary dyads or triads. Plutchik's eight basic emotions vary in polarity (the primary emotions are arranged in opposing pairs, e.g., joy vs. sadness), intensity (vertical levels of the cone; e.g., fear varies from terror to apprehension) and degree of similarity (adjacent pairs on the wheel reflect emotional similarity, e.g., anger is placed between disgust and anticipation) (Guerrero et al., 1998; Plutchik, 1980).

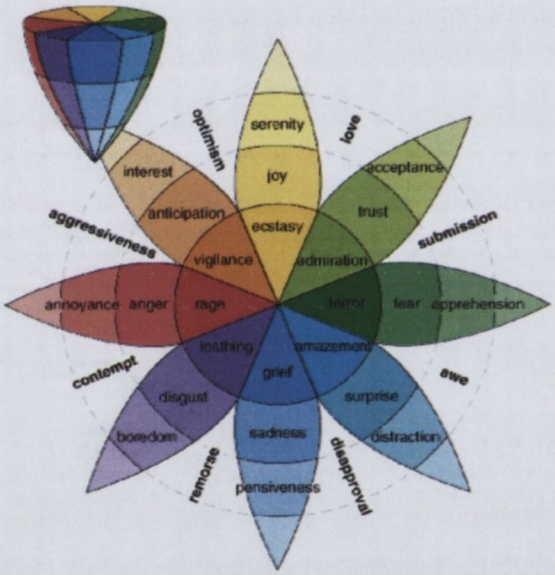


Figure 2.6. *Plutchik's multidimensional model of emotion (illustration retrieved from http://www.fractal.org/Bewustzijns-Besturings-Model/Nature-of-emotions.htm).*

THE PROTOTYPE APPROACH

Prototype approach models (also referred to in Scherer (2000c) as 'lexical models' or 'meaning oriented models') use the structure of the semantic fields of emotion terms to model the structure of emotion, e.g., Fitness and Fletcher (1993), Johnson-Laird and Oatley (1989), Ortony and Turner (1990), Shaver, Murdaya and Fraley (2001), Shaver et al. (1987). For example, Johnson-Laid and Oatley (1989) state that there is a small number of basic emotion modes roughly corresponding to happiness, sadness, fear, anger, and disgust that are referred to by a set of terms. Basic emotion modes have no internal structure, similarly, the terms designating these emotion modes are impossible to analyse semantically into anything more basic.

According to Shaver et al. (1987), emotion categories are formed as a result of repeated experiences and become organised around prototypes within an abstract-to-concrete hierarchy. In Fitness and Fletcher (1993), prototypes are defined as mental representations of categories of objects, events, and concepts like emotion. Each prototype comprises a set of organised features that characterise the typical example of a particular category or concept. In Fehr and Russell (1984), emotion's 'prototypicality' implies its 'goodness of example': emotion-denoting words can be ranked according to their being good or poor examples of emotion.

The emotions within each prototype family of emotion are organised according to two dimensions. The vertical dimension shows the hierarchical relationship among categories. The superordinate level (the most general level) is defined by the valence of the emotions; the basic (middle) level represents the core emotion within the category (e.g., love, joy, anger, sadness, fear, and surprise); the subordinate level consists of all other (non-basic) emotions that are related to the prototype. The horizontal dimension shows the relationship between members of the same category. The prototype approach considers thus both the contents of individual categories and the hierarchical relations among categories (e.g., loneliness is a type of sadness, which itself is a type of negative emotion) (Guerrero et al., 1998).

APPRAISAL THEORIES AND K. R. SCHERER'S COMPONENT PROCESS MODEL

Appraisal theories, e.g., Ellsworth (1991), Frijda (2005), Lazarus (1991), Scherer (1986) view emotions as resulting from a person's subjective evaluation (appraisal) of the significance of an emotion-eliciting object or event for the person's well being. The outcome of this appraisal process determines the resulting emotion reactions in terms of physiological arousal, expression, action tendencies, and feeling. Scherer (1999) points out that predictions formulated by appraisal theories concerning emotion elicitation and differentiation are relevant for full-blown emotion episodes (e.g., anger, fear, sadness, and so on) rather than for milder types of affective phenomena, such as moods.

The model that is particularly well-known in speech and emotion research is K. R. Scherer's component process theory, e.g., Scherer (1982a, 1984, 1988a) and also Johnstone and Scherer (2000), Sander, Grandjean and Scherer (2005), Scherer (1999, 2000c). The model treats emotion as a multicomponential phenomenon; the components of emotion process include 'the emotion reaction triad' (feeling, physiology and expression), as well as evaluation and action tendencies. Within the emotion episode, the changes in all these components are synchronised as the organism mobilises to cope with an important

event. The component process theory defines emotion as 'an episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism' (Scherer, 2004, p. 668). Emotional states are produced by the outcomes of a series of five stimulus evaluation checks (SECs): novelty check, intrinsic pleasantness check, goal/need significance check, coping potential check, norm/self compatibility check. The organism is seen as constantly scanning the environment evaluating and updating information on the basis of these five SECs; emotions result from specific combinations of outcomes of SECs. The effects of appraisal outcomes on vocal expression of emotion have been predicted in Scherer (1986) and supported by empirical research (e.g., Banse & Scherer, 1996; Juslin & Laukka, 2001, 2003). Scherer (1994) proposed the concept of 'modal emotions' to define a limited set of frequently encountered appraisal outcome patterns (combinations of evaluation checks results).

Scherer's component process model assumes that there are as many different emotional states as there are different patterns of appraisal results. The same event may provoke entirely dissimilar emotional reaction in different individuals or even result in no such reaction whatsoever (Scherer, 2004; Scherer, Banse, & Wallbott, 2001; Scherer, Wranik, Sangsue, Tran, & Scherer, 2004).

In summary, the different models outlined above can be applied to the research into vocal expression. They are not mutually exclusive, and each provides insights into how people categorise, express and perceive emotions. The auditory experiments reported in Chapters 4 and 5 of this thesis use discrete labels to define the anchors of the bipolar rating scales on which different synthesised voice quality stimuli were rated for the presence and strength of the affective colouring. These labels were purposefully selected to include 'basic' emotions (*sad, happy, scared*), milder affective states (*relaxed, bored*), attitudes and interpersonal stances (*formal, intimate, interested, apologetic*). As this thesis uses discrete emotional labels which are expected to be associated with particular voice qualities, it may appear that the theoretical stance adopted here is inclined towards the discrete emotions model. This does not mean, however, that dimensional approach cannot be applied here as each of the affective states referred to by a particular word can be represented in terms of the dimensional model of activation and valence. Discrete labels were chosen in part to ease the rating task for the participants: intuitively, this is how people describe the way they sound in everyday life.

2.3 Vocal expression of emotion

The following sections outline what is known to date about the role of voice quality in affect communication. Some major experimental findings as well as methodological problems relevant to the study of vocal communication of emotion will be outlined. It is worth noting that most studies combine encoding, decoding and acoustics, and that, for instance, where the major focus is the acoustics of emotional speech, the stages of emotion encoding as well as verification of the authenticity of expressed emotion through perception study is a standard procedure (e.g., Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005).

2.3.1 Collecting emotional speech data

Experimental studies of voice and emotion may be based on: (1) natural vocal expressions, e.g., Campbell (2004a), Campbell and Mokhtari (2003), Douglas-Cowie, Cowie and Schröder (2003), Douglas-Cowie, Campbell, Cowie and Roach (2003), Kappas and Poliakova (2008); see also review in Busso and Narayanan (2008); (2) induced emotion expressions, e.g., Bachorowski and Owren (1995), Johnstone (1996), Johnstone and Scherer (1999), Johnstone, van Reekum, Hird, Kirsner and Scherer (2005), Tolkmitt and Scherer (1986), Zara, Maffiolo, Martin and Devillers (2007), Caridakis et al. (2006); and (3) simulated/acted emotional expressions (used in the majority of studies), e.g., Banse and Scherer (1996), Bänziger (2004), Burkhardt et al. (2005), Carlson, Granström and Nord (1992), Juslin and Laukka (2001), Seppänen, Toivanen and Väyrynen (2003). Although acted speech is still used quite commonly (e.g., Bänziger & Scherer, 2007; Busso & Narayanan, 2008), the general trend in the field is to move towards the use of naturalistic speech corpora or induced emotional expressions.

Natural vocal expressions may be obtained from recordings of TV games or reality shows, e.g., the Belfast Naturalistic Emotion Database (Douglas-Cowie, Campbell et al., 2003) or speech recordings of everyday social interactions collected over a long period of time, e.g., the JST/CREST Expressive Speech Processing project (Campbell, 2005; Campbell & Mokhtari, 2003).

Such recordings have high ecological validity, but analysing such corpora presents numerous challenges. Firstly, they do not always have uniform quality and the recording conditions are such that they are unlikely to be suitable for certain types of acoustic analysis, such as inverse filtering (e.g., discussion in Cullen, Vaughan, Kousidis, & McAuley, 2008), for which very stringent recording conditions are needed if one wishes to obtain a detailed and accurate estimation of voice source characteristics. Burkhardt et al. (2005), Ní Chasaide and Gobl (2005) argue that natural data does not meet the

requirements necessary to ensure comparability of results such as a reasonable number of speakers and high quality laboratory conditions. The lack of control over the segmental content of the utterance could also be problematic as there can be large influences of segments on source characteristics.

Scherer et al. (2003) point out furthermore that it is not always easy to infer what emotion is being experienced by the speakers in naturalistic corpora. Batliner et al. (2003) comment that determining the beginning and end of a particular emotion episode presents the most important problem for the analysis of naturalistic emotional data. Another problem is that the material cannot be widely used due to copyright restrictions.

Vocal emotional expressions could be collected using various induction techniques, such as the Velten technique, the use of music and imagination, feedback mood induction procedure, social interaction, the use of mood-inducing films or slides, and more recently, specifically designed computer games, e.g., EmoTaboo (Zara et al., 2007) or sensitive artificial listener SAL (Douglas-Cowie, Cowie, Cox, Amir, & Heylen, 2008). The mood induction techniques are discussed in Martin (1990), Richell and Anderson (2004), Westermann et al. (1996). Westermann et al. (1996) compared the effectiveness and validity of 11 important mood induction procedures by meta-analysis of 250 studies. The presentation of a film or story turned out to be most effective in inducing both positive and negative mood states. The majority of mood induction techniques were found more effective for the induction of negative emotional states. The Velten technique (Velten, 1968) in which a number of statements describing either positive/negative self-evaluations or somatic states are presented to the subjects who are instructed to try to feel the mood described by these statements, is one of the most widely used (e.g., Barkhuysen, Krahmer, & Swerts, 2007; Wilting, Krahmer, & Swerts, 2006).

Mood induction techniques have the advantage of providing the samples of high ecological validity from a large number of speakers allowing for control over verbal and emotional content. Scherer (2003) points out, however, that this method produces relatively low in intensity and not clearly differentiated emotional expressions. Frequently arising ethics-related problems is another downside of induction techniques. Induction methods have been used, for example, in Aubergé and Cathiard (2003), Audibert, Aubergé and Rillard (2005), J. Barrett and Paus (2002), Batliner et al. (2003), Batliner, Steidl and Nöth (2008), Rossato et al. (2004), Caridakis et al. (2006), Zara et al. (2007).

Simulated emotional expressions are obtained by asking professional, lay actors or indeed non-actors to portray certain emotions, based on emotion labels or more detailed descriptions. The actors may be allowed to freely use self-induction techniques, as in

Abelin (2004), Airas and Alku (2004), Burkhardt et al. (2005), Sawamura et al. (2007), or use scenarios, e.g., Scherer et al. (2001), Scherer, Banse, Wallbott and Goldbeck (1991), Schröder (1999), Sobin and Alpert (1999), Williams and Stevens (1972) or/and the guidance of professional director (Banse & Scherer, 1996; Bänziger & Scherer, 2007; Scherer et al., 2001). Different acting styles and methods to ensure consistent elicitation are discussed in Enos and Hirshberg (2006). In certain cases, actors are asked to imitate real-life speech with the same lexical and emotional content in laboratory settings, e.g., (Williams & Stevens, 1972) or more recently, the Belfast Structured database (Douglas-Cowie et al., 2006).

According to Scherer (2003), simulated data can produce much more intense prototypical expressions compared to induced emotions and allows experimental control over the lexical content of the utterances and emotions expressed as well as cross-speaker comparison of results. However, the approach has been criticised (Bachorowski & Owren, 2003; Russell, Bachorowski, & Fernández-Dols, 2003) and also discussion in Bänziger and Scherer (2007) mostly on the grounds that portrayals may reflect exaggerated culture-specific stereotypes rather than genuine real life expressions. The perceived genuineness of such feigned emotions depends highly on the skills of the actor and the range of expressive tools in his/her repertoire. This may result in inconsistencies and idiosyncratic preferences in expression, and so conclusions based on the analysis of acted data are less generalisable. Acted expressions are controlled, and thus subtle involuntary physiological changes related to sympathetic arousal such as breathing patterns and laryngeal tension etc. that are common in spontaneous reactions are unlikely to be present. Acted expressions in general are viewed as a simplified representation, while in real life emotions are expressed in interaction and context, as complex blends rather than pure full-blown emotion episodes and via a combination of multiple channels (voice, face, gesture, posture)(Douglas-Cowie et al., 2005). Thus recognition systems trained on acted speech show poor performance on natural data (Batliner, Fischer, Huber, Spilker, & Nöth, 2000)[6].

The use of acted speech is therefore limited by the elicitation procedures, skills of the actors, the accuracy/precision of specification /description of the affective states that are of interest (broad basic emotion terms will result in variability of expression); the specification of context (a scenario describing the situation in which the emotion is

[6] Discussion of emotional speech recognition techniques is beyond the scope of this thesis; review of such techniques in the study of emotionally loaded speech as well as the most frequently used acoustic features is given, for example, in Ververidis and Kotropoulos (2006).

elicited); the possibility of interaction with partners (more naturalistic as emotions are communicated) and the use of other expressive modalities.

In defence to this approach, Scherer (2003, p. 144) points out that 'all publicly observable expressions are to some extent 'portrayals', given the social constraints on expression and unconscious tendencies toward self-presentation'. The fact that vocal portrayals are reliably recognised by listeners in perception experiments proves that they are representative of naturally occurring emotional expressions. Bänziger and Scherer (2007) argue that when well designed, emotion portrayals are a viable and useful research methodology.

Acted or portrayed affect should be presented as such, and the extent to which findings based on such samples represent the real world of emotion/affect expression has to remain an open question, to be eventually verified by some other approach. In this thesis, an analysis of a fragment of acted or portrayed affect is analysed in minute detail for the voice source correlates that pertain to each affect (Chapter 6). The fact that the desired states (*bored, sad, happy, surprised, angry* and *neutral*) were indeed communicated, although admittedly with varying degree of success, was confirmed in the course of a listening test. In support of the argument expressed by Scherer, all portrayals are in some way representative of the way people use voice in real world. We are not claiming that this is necessarily indicative of 'true affect'. Rather, these may be indicative of the kind of portrayals that pepper everyday conversations. Finally, they are likely to be quite suitable for implementation in certain restrictive contexts, e.g., synthetic narration of children's fiction. We do not wish to claim that these productions are anything but portrayed emotions, with all concomitant limitations and implications for the interpretation of results. The validity of the short experiment reported in Chapter 6 is consequently constrained.

The extent to which portrayals differ from natural voice expression has been studied in Audibert, Aubergé and Rilliard (2008), Barkhuysen et al. (2007), Mathon and de Abreu (2007), Shahid, Krahmer and Swerts (2008), Wilting et al. (2006). For example, in Barkhuysen et al. (2007), Czech participants rated the perceived emotional state, acted or real, of Dutch speakers. It was found that acted emotional speech leads to significantly more extreme perceived emotion scores than non-acted emotional speech, and that the difference between acted and real emotional speech is larger for the negative than for the positive conditions.

The utterances used in portrayed emotion expression are usually content-neutral, although there is growing awareness that this approach is probably too simplistic (Douglas-Cowie et al., 2005). Generally, alphabet or numbers, or repetitions of a single word (e.g.,

name), as in Abelin (2004), Hammerschmidt and Jürgens (2007), Leinonen, Hiltunen, Linnankoski and Laakso (1997), Linnankoski, Leinonen, Vihla, Laakso and Carlson (2005), Sawamura et al. (2007), or specially constructed nonsense sentences, e.g., Banse and Scherer (1996), Bänziger (2004), Bänziger and Scherer (2007), Breitenstein, van Lancker and Daum (2001), have been used. For example, in Scherer et al. (2001) 'meaningless multilanguage sentences' were constructed by selecting two syllables from six European languages, German, English, French, Italian, Spanish, Danish, and randomly arranging them into several seven-syllable sentences. A group of expert listeners selected two sentences that were most appropriate in terms of the ease of articulation, language neutrality and status as sentences, for the use in a cross-language perception test. It could be argued, however, that speech-like segments in such presumably content free voice samples still carry information of linguistic similarity/dissimilarity relative to the language of decoders (e.g., complexity of consonant clusters permitted by the phonological structure or durational characteristics of vowels) and could potentially influence the listeners' judgment.

Burkhardt et al. (2005) argue that the use of everyday-life sentences is preferable as quasi-spontaneous emotional expressions are more difficult for the actors to produce using nonsense material. Furthermore, it is virtually impossible to construct sentences completely devoid of the effects of segmental and suprasegmental aspects of language.

2.3.2 Perceptual approaches

Perception studies in vocal expression and communication of emotion have been mainly of two kinds. The first – emotion recognition studies – are focused on the ability of listeners to recognise emotions expressed in voice. The second – 'cue utilisation' studies – consider the relative importance of vocal cues that the listeners use to infer emotional content in voice. Recent reviews of methods in perception studies are found in Elfenbein and Ambady (2002b), Johnstone and Scherer (2000), Juslin and Laukka (2001, 2003), Juslin and Scherer (2005), Kappas, Hess and Scherer (1991), Ní Chasaide and Gobl (2005), Scherer (2003), Scherer et al. (2003).

EMOTION RECOGNITION STUDIES

As a rule, the listeners are presented with recorded content-free speech samples of (usually acted) emotional expressions and are asked to select the perceived emotion from a list of emotion labels, and sometimes also to rate the intensity of expressed emotion. The accuracy of recognition usually reported for acted emotions is about 60 per cent, which is

lower compared to the studies of recognition of emotion from face (75 per cent accuracy) (Banse & Scherer, 1996; Scherer et al., 2003). Only a small number of emotions have been studied (e.g., the 'Big Six') with only one positive emotion (joy) included in the list of emotion categories. Activation-related or intensity-related variants of the same emotion have rarely been taken into consideration[7]. Most studies have used forced-choice listening task (when only the labels for 'correct' emotion categories are included in the answer sheet). The recognition/attribution is considered correct if it matches the emotion purposefully expressed by an actor. Therefore, such studies have based their results on discrimination rather than recognition. Russell (1994) argues that forced-choice studies present affective labels as mutually exclusive thus preparing participants to interpret stimuli as expressions of emotions.

This inflation of the accuracy of emotion recognition could be corrected by using statistical procedures, such as Wagner's 'unbiased hit rate' (Wagner, 1993). This has been suggested by Scherer (2003) as well as other researchers (Elfenbein & Ambady, 2002b; Elfenbein, Ambady, Mandal, Harizuka, & Kumar, 2002). Furthermore, labels 'no emotion' or 'none of the above' could be added to the list of alternatives, or the listeners could be allowed to freely label the perceived emotion content of the speech samples (e.g., Abelin & Allwood, 2000; Greasly et al., 2000). However, free labelling is often difficult to interpret, and is particularly problematic for cross-language studies as there may be no direct correspondence between emotion-denoting words in different languages. An interesting approach in an effort to avoid the explicit use of emotion terms in the response sheets was adopted in a cross-cultural study reported in Burkhardt et al. (2006). The authors coded the emotion-related states in phrases that might have been uttered in emotionally loaded situations. For example, the phrase 'Congratulations, you've just won the lottery!' was expected to be uttered with joy, whereas the emotion expected to colour the phrase 'This is the five thousand three hundred second status report. All systems are up and running' was boredom. The participants in the listening tests had to answer one question: 'Was the sentence uttered in an appropriate way?' and rate the appropriateness of the manner of speaking for the meaning of the sentence.

Although frequently criticised, the forced choice paradigm has the advantage of being less tiresome for participants (compared to dimensional ratings) as well as easier to interpret and less time consuming for the experimenter (compared to free labelling). Elfenbein et al. (2002) report that forced choice design studies and those using alternative

[7] In Juslin and Laukka (2001), listeners decoded actors' portrayals of happiness, sadness, anger, fear, and disgust with weak and strong emotion intensity. The results showed that portrayals with strong emotion intensity yielded higher decoding accuracy than portrayals with weak intensity.

methods yield largely similar results. Cowie and Cornelius (2003) also acknowledge the advantages of the forced choice methods provided the lists of options exceed the six basic emotion categories.

DIFFICULTIES WITH CROSS-CULTURAL STUDIES

When testing cross-linguistic perception, the emotion labels used as response choices in perception studies are usually selected by European or North American researchers and subsequently translated into the languages of interest for cross-cultural studies. However, these original labels do not necessarily correspond to emotion constructs in other cultures (Ogarkova, Borgeaud, & Scherer, 2009).

Russell (1991) argues that, since emotion words vary with culture, it is possible that persons from different cultures might encode, respond to and remember emotions in different ways. Languages differ in the number of words they provide to categorise emotions (some languages do not have equivalents for emotion-denoting words in English and vice versa). Oatley et al. (2006) give an example of such variability: researchers have identified 2,000 emotion-related words in English, 750 in Taiwanese, 58 in Ifaluk of Polynesia and 8 of the Chewong of Malasia (p. 183). Furthermore, cultures vary in which states they represent with emotion terms. E.g., the distinction between shame and embarrassment is not made by the Japanese and Indonesians. A list of emotion categories specified by English words and then translated may not necessarily contain the word that the subjects would find most appropriate.

Russell (1991) recommends free labelling as the most desirable approach in the cross cultural studies. This, however, may not be particularly helpful, as the researcher (a speaker of language A) will still need to interpret the labels obtained as a result of free labelling (by the subjects – speakers of language B). Other methods, e.g., those recommended in Brislin (1980), Streiner and Norman (2008) such as team translation, when a number of bilingual speakers translate the affective terms from language A into language B, discuss all possible variants and reach consensus. These translations are then translated back into language A by a different panel of bilinguals and subsequent discrepancies in the resulting translation are carefully considered.

It is clear, therefore, that response choices used in the perception tests are constrained. The shorter the list of response choices, the broader the possible interpretation of the emotive words by the listener judges and the broader the set of emotions (the whole family) that one label can mean. The shorter and the less specific the description is, the more it is open to interpretation to the listeners. Furthermore, if the listeners are from

different cultures, the quality of translation, the difference in the semantic fields of emotion denoting words and familiarity with the terms will directly reflect on the recognition rates.

Perception studies report emotion-specific recognition accuracy, and this accuracy is different for facial and vocal expression. For example, joy is perfectly recognised in the face, but not in the voice; sadness and anger are generally best recognised in the voice, followed by fear; disgust in the voice is recognised barely above chance level yet it is reliably recognised from the face (Scherer, 2003). Meta analysis conducted by Juslin and Laukka (2003) based on 39 studies of decoding accuracy of emotion from voice reports that sadness and anger portrayals are decoded most accurately, followed by fear and happiness; the emotion recognised with lowest accuracy was tenderness. It has been suggested that the differences between recognition accuracy of emotions from face and voice can be explained by evolutionary pressure (relative frequency of occurrence of different expressed emotions as well as relative adaptive importance of recognising such emotion), e.g., detecting fear/anger is more important for survival than recognising happiness (Johnstone & Scherer, 2000).

Patterns of confusion in the emotion recognition studies are not random: usually, emotions similar in terms of quality, intensity, or valence are more likely to be confused (Johnstone & Scherer, 2000; Scherer, 2003). For example, in Elfenbein et al. (2002), hot anger has been reported to be confused consistently only with cold anger and contempt.

Traditional views that emotion expression/communication process involves the stages of encoding and decoding have been challenged in Russell et al. (2003). They suggest that emotions are not encoded and 'broadcast' by the speaker and subsequently decoded by the listener. Rather, emotion expressions have evolved to influence the listener in ways beneficial to the sender. Vocal stimuli can capture the receivers' attention and alter their affective state without any messages being decoded. The authors suggest that rather than judging whether emotions are recognised correctly or incorrectly, the focus in emotion expression and communication studies needs to be shifted to the (separate) description of the processes involving receiver's response (which is not limited to emotion attribution or might not include such attribution at all) and sender's production of emotion expressions.

CUE UTILISATION STUDIES

Perception 'cue utilisation' studies aim at establishing the relative contribution of various voice cues or cue combinations in emotion signalling, e.g., Breitenstein et al. (2001), Burkhardt and Sendlmeier (2000), Carlson et al. (1992), de Gelder and Vroomen (2000), Gobl and Ní Chasaide (2003b), Klasmeyer (1997), Ladd, Silverman, Tolkmitt,

Bergmann and Scherer (1985), Protopapas and Lieberman (1997), Scherer, Ladd and Silverman (1984), Mozziconacci (1998), Teshigawara (2004). The listeners are presented with stimuli in which certain acoustic cues have been systematically removed or manipulated through masking (e.g., band-pass filtering) or synthesis; the relative importance of particular acoustic cues is inferred from the listeners' ratings of such stimuli. Elimination of a prosodic feature that is relatively unimportant for emotion signalling is assumed to result in no significant decrease in affective rating.

For example, a series of studies reported in Scherer et al. (1984), tested the effects of f_0 range and f_0 contour and voice quality on affective ratings using several masking techniques. Random splicing destroyed the f_0 contour shape and the verbal content, but kept the global f_0 level and range as well as the voice quality information. Low-pass filtering kept the f_0 contour but removed the voice quality cues and the verbal content. The results suggested that f_0 contour was of relatively little importance, relative to the voice quality cues, for the transmission of affective content. In a series of listening tests, Mozziconacci (1998) used utterances resynthesised based on the neutral rendition but with experimentally obtained affect-related f_0 contours in a series of listening tests. She reported a substantial drop in emotion recognition accuracy of the utterances from which affect-related voice source information had been removed. The results support those of Scherer et al. (1984).

Breitenstein et al. (2001) compared the perception of simulated expressions of happy, sad, angry, frightened, and neutral states in which f_0 and speech rate cues were systematically varied. The forced choice based listening tests showed that of those two variables, speech rate was the most potent cue, with f_0 variation contributing less to emotion attribution. For example, slow rate was associated with sadness, whereas fast rate was reliably associated with angry or frightened affects.

Audibert et al. (2005) used different types of resynthesis of monosyllabic words from a database of acted emotional speech to identify the relative weight of various voice cues in communication of anxiety, disappointment, disgust, disquiet, joy, resignation, satisfaction and sadness. Resynthesis involved different combinations of parameters related to f_0, intensity and voice quality, e.g., full resynthesis; neutral expression + stylised emotional f_0 contour (f_0 only); neutral + stylised intensity contour (intensity only); f_0 and intensity (both contours applied) etc. The results of the listening tests showed, for example, that expressions of joy obtained the highest scores for the f_0 only and f_0 and intensity stimuli, whereas expressions of disgust, sadness and fear were attributed highest ratings in voice quality and duration stimuli. No one parameter was found to reliably carry the whole

emotion information, with f_0 contours revealing more information on positive expressions and voice quality and duration conveying more information on negative expressions.

Burkhardt et al. (2006) used synthesised stimuli to study the effects of systematic manipulation of pitch range, duration, and jitter on emotional speech perception by the speakers of French, German, Greek and Turkish. The results showed language-specific reactions to manipulation of the three parameters.

Waaramaa et al. (2006) used resynthesis to manipulate F3 (the third formant) in the [a:] vowel extracted from acted expressions of sadness, tenderness and anger. The results show that F3 may affect perception of valence: raised F3 was associated with positive valence more often than the samples with the original F3, with the F3 lowered by 30 per cent or with F3 removed.

Scherer (2003), Scherer et al. (2003) suggest that perception studies need to examine psychoacoustic properties of affective speech (e.g., perceived loudness, perceived pitch, perceived rhythm etc.) as well as the role of transmission of the voice signals: (1) through space and electronic channels and (2) through the human hearing mechanism. Such studies are, he believes, likely to provide parameters for emotion differentiation that have been missed in the studies of classical acoustic cues.

The perceived features of emotional speech and their correlation with acoustic parameters were studied in Bänziger and Scherer (2003). The listeners judged acted expressions of four emotions, each with two activation levels, on a scale[8] presented on a computer screen together with two recordings representing extreme exemplars of the cue measured, e.g., 'pitch' (low ↔ high); 'intensity' (weak ↔ strong); 'intonation' (monotonous ↔ modulated); 'speed' (slow ↔ fast); 'articulation' (bad ↔ good); 'instability' (steady ↔ shaky); 'roughness' (not rough ↔ rough); 'sharpness' (not sharp ↔ sharp). The listeners were asked to adjust the cursor along the scale according to the degree of the feature presence. Since the listeners were provided with two auditory benchmarks, high interrater agreement was achieved in the study. Average intensity was found to be the best predictor of perceived intensity and 'sharpness', total duration - the best predictor of perceived speed; f_0 range - the best predictor of perceived intonation and f_0 minimum is the best predictor of perceived pitch.

Methodological problems related to the perception studies concern the use of rating scales, e.g., Gannon and Ostrom (1996), Granqvist (1996), Russell and Carroll (1999), the problems of listeners agreement, e.g., Kreiman and Gerratt (1998, 2000b), Kreiman, Gerratt and Ito (2007), Shrivastav (2003, 2006), Streiner and Norman (2008).

[8] The presentation of the scale was based on Granqvist (1996).

2.3.3 Vocal expression of emotion: acoustic measures

Expression of emotion in speech has been studied extensively through other measurements related to the acoustic analysis of the speech waveform (and to a much lesser degree through physiological measures), for reviews, see Banse and Scherer (1996), Cowie and Cornelius (2003), Frick (1985), Gobl and Ní Chasaide (2003b), Johnstone and Scherer (2000), Johnstone et al. (2007), Juslin and Scherer (2005), Laukka (2008), Murray and Arnott (1993), Scherer (1986, 2003), Scherer et al. (2003), Toivanen, Seppänen and Väyrynen (2009).

The most commonly studied vocal parameters are those related to f_0, duration, intensity and voice quality. Most of the above studies have tended to focus on f_0, intensity and durational factors, and often only large global statements about these acoustic parameters are reported. These are indicative but do not provide insight into the dynamics of the voice source.

Note that most studies give broad, averaged or global results, e.g., greater level and range of f_0 for high activation emotions (Banse & Scherer, 1996) or a global NAQ values (Airas & Alku, 2004; Campbell, 2004a; Toivanen et al., 2006). Fónagy (2001, p. 91) argues that emotion differentiation based on the mean values of acoustic measurements is doomed to failure as the essentially dynamic nature of emotion expression is not adequately represented by such measures. Capturing dynamics of emotionally coloured speech is therefore another direction that research needs to focus on.

The focus of this study is on the voice quality. As outlined above, voice quality is difficult to measure directly due to a number of reasons including very stringent requirements of the recording, the somewhat limited accuracy and robustness of most available analytic techniques, complexity of data interpretation etc. Some of the findings pertaining to the study of emotion in speech are outlined below.

VOICE SOURCE ANALYSES

Analysis of the voice source with regard to emotion-specific patterns is expected to contribute significantly to the existing acoustic measurements described in section 2.3.3 above, in particular where emotion differentiation is of interest. However, research of the role of the voice source in emotion expression is plagued with a number of methodological problems, outlined, for example, in Gobl and Ní Chasaide (2003b), Ní Chasaide and Gobl (2005). The limited availability of appropriate analytical tools and the rigid requirements to the quality of the speech signal to be analysed are among such problems. Furthermore, in the analysis of voice source it is problematic to divorce emotion-related source variation and other determinants such as the speaker's inherent voice quality and dynamic variations

introduced by the segmental and suprasegmental characteristics of an utterance, e.g., segment boundaries or syllable prominence. These considerations make the generalisations based on the automatic analysis of relatively large amount of emotional speech data somewhat questionable.

Several studies of vocal expression of emotion have used Normalised Amplitude Quotient (NAQ), e.g., Airas and Alku (2004), Campbell (2004a), Campbell and Mokhtari (2003), Rossato et al. (2004), Toivanen et al. (2006), Laukkanen et al. (2008), Waaramaa, Laukkanen, Alku and Väyrynen (2008). For example, Campbell and Mokhtari (2003) used NAQ for analysis of voice quality in terms of breathy - pressed phonation continuum in a large conversational speech corpus in Japanese (female speaker). They found that voice quality as reflected by the NAQ value varied consistently to signal paralinguistic information depending on such factors as interlocutor (e.g., family, friends, others, self-directed), speaking-style (polite, friendly, casual), and speech-act ('giving information', 'exclamations', 'requesting information', 'muttering', and 'requesting repeats'). The degree of breathiness was reported to rise (high NAQ) in family members directed speech compared to friends directed; self-directed speech exhibits the lowest NAQ values. The NAQ was therefore reported to correspond with the 'degree of care' in speech and to vary independently of f_0. In Airas and Alku (2004), Laukkanen et al. (2008) NAQ was used in the analysis of short [a:] vowel segments extracted from emotion portrayals of neutral, sadness, joy, anger, tenderness, as an indicator of emotion differentiation. The results showed significant differences among most of the studied emotions as well as clear gender differences. For example, relative to neutral, anger showed lower mean NAQ value whereas joy, sadness and tenderness all exhibited higher mean NAQ values, indicating 'smoother glottal pulse' compared to neutral. The considerably higher standard deviation of NAQ values for joy, sadness and tenderness compared to neutral were reported as indicative of larger variation in the voice quality within these emotions. The NAQ values, ranked from lowest to highest, were found for angry, neutral, joy, sadness and tenderness respectively. In Laukkanen et al. (2008), NAQ was shown to reflect activation dimension in emotional speech.

The shape of the glottal pulse has been suggested as a variable that carries information on the emotional state of the speaker. Few studies, however, have considered the shape of the glottal pulse with particular regard to emotional speech or emotion differentiation. For example, Cummings and Clements (1995a) examined the glottal

waveform for 11 different speaking styles (normal, angry, 50 per cent tasking[9], 70 per cent tasking, clear, fast, Lombard, loud, question, slow, and soft) in order to establish a set of glottal waveform shape parameters that exclusively describes each of these styles. The data (non-nasalized vowels from the words *fix, six* and *destination* in the production of two speakers of General American) was extracted from the Lincoln Labs Multi-Style Speech Database. The glottal waveforms were obtained by an inverse filtering method and subsequently parameterised in terms of six wave shape features: closing slope, opening slope, closing duration, closed duration, opening duration, and top duration. The data analysis showed that the glottal waveform characterising each speaking style was significantly different from those of all other styles; these differences were particularly prominent for angry, loud, and soft styles. The authors conclude that the shape of the glottal waveform is reflective of the speaking style and varies as its function. Similar findings were reported by Laukkanen, Vilkman, Alku and Oksanen (1996) who used the Iterative Adaptive Inverse Filtering method (IAIF) (Alku, 1992) to analyse portrayals of surprised, sad, enthusiastic, angry and neutral on a nonsense utterance 'paappa p*aa*ppa paappa'. They conclude that the glottal waveform has significance for the identification of the emotional states. In a further study, Laukkanen et al. (2008) report that the type of glottal source is used independently of f_0 level, intensity and duration. Klasmeyer (1997) looked at the glottal pulse shapes (obtained by inverse filtering) and the spectral tilt as the acoustic correlates of the vocal effort and phonation types using a short utterance on which a number of emotions were simulated. This study showed that the glottal pulse shape, along with other acoustic parameters (spectral tilt, spectral distribution of turbulent signal components and voicing irregularities) is useful for differentiating between the emotions happiness, sadness, anger, fear and boredom.

Analysis of the glottal waveform features, such as the lengths of the opening phase (OP), closing phase (CP), closed phase (C), open phase (O), and total cycle (TC) has been applied in the studies of emotion-related disorders such as clinical depression and are discussed, for example, in Moore II, Clements, Peifer and Weisser (2003, 2008). Murphy and Laukkanen (2009) recently reported a preliminary study of glottal characteristics of acted affective states sad, tender, joy, anger and neutral using the electroglottogram signal and its first and second derivatives. The results based on the measures of closed quotient and rate of vocal fold contact at closure and opening for the sustained vowel [a] suggest emotion-specific differences. Rank-ordering the values of contacting quotient CtQ (contact

[9] 50% tasking and 70% tasking speech is produced while the speaker is under a specified level of workload; Lombard speech is produced while the speaker is subjected to 90 dB background noise.

time of the vocal folds divided by cycle duration) was angry > joy > neutral > sad > tender. This suggests a more pressed phonation for anger and a more lax configuration for sad and tender (as expected). Contacting quotient appears to separate joy from neutral better than contact quotient. Rank ordering speed quotient and acceleration of contact at closure values showed the lowest value for joy, followed by neutral > angry > sad > tender.

In general, these studies underline how limited is our understanding of the use of the voice source in affect signalling. However, there are certain clear findings that can help advance the enterprise of developing more flexible voice source models for application in synthesis, a hope expressed also in Burkhardt (2009), Burkhardt and Stegmann (2009).

The studies reported in thesis first and foremost aim at getting a better understanding of how voce quality maps to affect and whether this mapping is similar across the languages studied or shows prominent cross-language variation. There are potential downstream technological applications of the kinds of results that emerge from this study. One concerns the area of affective synthesis[10] (discussed, for example, in Taylor (2009), Burkhardt and Stegmann (2009), Bulut, Narayanan and Johnson (2004), Campbell (2008), Schröder (2001, 2004b, 2009)), at least if one is using a parametric approach, either using the traditional formant-based synthesis (Burkhardt, 2009) or the now rapidly developing statistical machine learning techniques such as HMM and voice conversion techniques, where, in principle, source and filter can be separately controlled (Cabral & Oliveira, 2006).

Findings such as presented here, however exploratory, may be of use to those working in the field of expressive speech synthesis. Particularly insofar as certain voice qualities can signal radically different affects for different language groups, some of the results would appear highly relevant. See, for example, section 4.4 in Chapter 4, where stimuli associated with intimacy by Japanese and Spanish speakers would appear formal to the English or Russian speaker's ear. Clearly, for anyone wanting to build a synthesiser with a friendly intimate voice would do well to take account of such kinds of cross-language differences as may arise, to ensure that the voice generated is indeed appropriate to that affect.

The principal goal of the present study is, however, to gain more understanding of how the voice is used to signal affect, and how languages may differ in their exploitation of voice quality. Potential technological applications are not a direct concern, and will not be further discussed here.

---

[10] Samples of currently available expressive synthesis collected by Felix Burkhardt are presented on his website http://emosamples.syntheticspeech.de/index.html. The description of the main systems used for synthesis can be found on http://www.synsig.org/index.php/Software.

INDIRECT VOICE SOURCE MEASURES

The data on voice quality can also be collected from the indirect measures related to the spectral characteristics of the speech signal. For voice quality in particular, the most commonly calculated indirect measures are: the amount of high frequency energy measured as the relative proportion of total acoustic energy above, versus below, a certain cut-off frequency, obtained by measuring LTAS (this also is considered a measure of spectral slope); formant frequencies and bandwidths; precision of articulation measured as the deviation of formant frequencies, usually that of F1, from the neutral baseline; the shape of the glottal waveform (Juslin & Scherer, 2005). Some of the studies that used these measures will be discussed below.

OTHER ACOUSTIC MEASURES

Apart from the measures related to voice quality characteristics, the most commonly studied vocal parameters are those related to f_0, duration and intensity, and it is well established that these prosodic characteristics contribute in an important way to the expression of affect. The measures are usually global and include f_0 mean and standard deviation, description of (often stylised) f_0 contour; jitter; intensity, 'attack' (measured as the rise time of amplitude of voiced segments); shimmer; speech rate (e.g., measured as the number of syllables per second); pauses; rhythm (Juslin & Scherer, 2005).

The findings for particular emotions reported in the literature are summarised in Table 2.4.

THE ROLE OF F_0

The role of f_0[11] in affect signalling has been studied extensively and f_0 variation has long been considered as a primary carrier of affective information. Some of the issues related to f_0 are pertinent to the experiment reported in Chapter 4 and will be discussed below.

Global f_0 level forms a part of the cross-species biological 'frequency code' (Ohala, 1984, 1994, 1996) that suggests that pitch is associated with the size of the vocaliser and thus with the amount of danger/threat it might present. Larger and stronger animals are more likely to produce low-pitched sounds, while high-pitched signals suggest a smaller size animal. Low pitch signals have therefore developed to signal aggression and dominance, and could be predicted in the expressions of disgust and contempt whereas high-pitched sounds show lack of aggression. The 'frequency code' is further evident in

[11] f_0 variation as an acoustic variable must be distinguished from intonation as a system of linguistically relevant categorical distinctions, as argued in Ladd, Scherer and Silverman (1986).

cross-culturally similar uses of f_0 to signal affect: e.g., deference, politeness, submission, lack of confidence are signalled by high and/or rising f_0 whereas assertiveness, confidence, authority, aggression, threat are conveyed by low/falling f_0 (Ohala, 1984, 1994)

Scherer and colleagues (Ladd et al., 1985; Scherer et al., 1984) proposed that affect is signalled differently by two types of cues: (1) 'continuous' acoustic parameters that are directly correlated with variations in the affective message, in particular with the physiological arousal of the speaker (e.g., the higher the average f_0 of an utterance the higher the level of arousal of the speaker), and that act independently of the linguistic meaning - 'covariation model' and (2) linguistically relevant acoustic features (categorical linguistic structures) that can only be interpreted in combination with verbal content (e.g., a falling contour is neutral in a wh-question but sounds challenging or aggressive in yes-no question) - 'configuration model'. Configuration model is more suitable for the description of complex affective states that involve a substantial cognitive component (e.g., attitudes), and the covariation model better explains biologically based emotions (Grandjean et al., 2006; Kappas et al., 1991).

These two approaches, also referred to as 'continuous' and 'cue-configurational' (e.g., Bachorowski & Owren, 2008), reflect the two main strands of research in the acoustics of affective speech. In relation to the study of f_0, the continuous approach would suggest that f_0 level and range vary continuously with physiological arousal, while the 'cue-configurational' one would suggest existence of emotion-specific f_0 contours.

Conflicting results have been reported as regards the ability of specific shapes of f_0 contour to carry affective information. Several authors reported existence of such contours (Fairbanks & Pronovost, 1939; Fónagy & Magdics, 1963; Williams & Stevens, 1972). In a recent study, Audibert and Aubergé (2007) found conclusively that affective information in expressions of joy and satisfaction was mainly carried by f_0 contours.

The shape of intonation contour has long been related to the attitudinal function of intonation. O'Connor and Arnold (1973) comprehensively explored this function of English intonation, grouping particular pitch patterns within an intonation phrase ('tunes'), into ten 'Tone Groups' according to the attitudinal meaning the pattern conveys in association with five main sentence types (statement, yes-no question, wh-question, command and interjection) and the meaning this tune conveys in spoken communication. For example, statements (declaratives) intended to correct or contradict the information provided by one of the interlocutors may sound concerned, reproachful or hurt when produced with the Switchback (fall-rise nuclear tone), and resentful when produced with the Take-Off (a low rise nuclear tone combined with a low head). Attitudinal function of

intonation (in English) is further discussed in a more recent account of English intonation (Wells, 2006). Here, for example, the same (in terms of segmental components) command - 'Now move along, please' may sound firm and affirmative (with the Low Fall), urgent and warning (with the Rise-Fall) or routine and friendly (with the Low Rise) (Wells, 2006, p.62). This pragmatics-based description does not explain, however, whether and to what extent pitch contour on its own, that is, without the verbal content, is responsible for signalling attitude. Wichmann (2000) argues the attitudinal meaning of a prosodic contour can only be inferred from the context (and varies as its function) and often arises from a mismatch between the prosody and the content/context of the message. Similar is suggested in Bänziger, Morel and Scherer (2003).

Uldall (1964) conducted perceptual study of the emotional and attitudinal meaning of intonation using Osgood's semantic differential dimensions (bored-interested, polite-rude, deferential-arrogant, tense-relaxed etc.). Sentences representing four main communicative types (statement, yes/no question, wh-question, command) were combined with 16 synthetically manipulated intonation contours varying in range, pitch level at the end of the contour, and the shape (one direction/with a change of direction). The results showed three main underlying 'dimensions of meaning' conveyed by these contours: pleasant/unpleasant; authoritative/submissive; strong/weak. The emotional meaning of different contours was found to depend on the communicative type of the carrier sentence. The same contour was rated as corresponding to a 'weak' emotional intensity when applied to the statement and the two types of questions but as corresponding to a 'strong' emotional intensity when applied to the command utterance. The variability in emotional meaning was also linked to the contour dynamics: the less dynamically varied the contour is the more variable it is in meaning. Scherer et al. (1984) reported similar effect of specific combinations of f_0 contours (e.g., final rise or fall) and linguistic category of the utterance (type of question) on affective colouring of the utterance.

The studies of f_0 contours in content-free sentences report no evidence for affect-related f_0 contours. Mozziconacci (1998, 2002) found no one-to-one correspondence between emotion and f_0 contour: one f_0 pattern can be used in the expression of different emotions, but with differences in pitch range and pitch level. Conversely, structurally distinct f_0 patterns can be relevant for emotions expressed with the same pitch range and pitch level. The most frequent pattern in the dataset used in her study was '1&A'[12] (rise-fall on a single accented syllable). The author concludes that this pattern is the most likely

[12] Mozziconacci used the Dutch grammar of intonation ('t Hart, Collier, & Cohen, 1990) in which numbers are used for rises and letters for falls.

default pattern of pitch movement and the most suited pattern to control for pitch variability.

Similar results are reported in Bänziger (2004), Bänziger et al. (2003), Bänziger and Scherer (2005) who analysed a large corpus of acted emotional speech (Banse and Scherer, 1996) to establish whether emotions (anger, fear, sadness and joy; each with two levels of activation) were expressed with particular f_0 contours. Results show little evidence for qualitatively different contour shapes for different emotions. Global measures (mean level and range of f_0) were sufficient to show strong differentiation of emotions and varied strongly with the degree of activation. There was some evidence for the importance of f_0 dynamics in the expression, e.g., the steepness of the rising and falling slopes and the relative height of accent peaks tend to be different for high- vs. low-aroused emotions. The importance of the relative height of the final f_0 peak and its dynamics for emotion differentiation is also reported in Mozziconacci (1998).

There is some evidence, however, that f_0 contours can carry emotional meaning independently of linguistic content. The results of the experiments using various masking procedures reported in Ladd et al. (1985), Scherer et al. (1984) showed that all nonverbal aspects of the utterances tested - f_0 range and contour, and voice quality in particular, carried the emotional meaning independently of the content. However, these experiments also suggest that the shape of f_0 contour is of relatively less importance compared to the global f_0 level and spectral characteristics of voice quality. Stimuli generated by random-splicing (which destroys f_0 contour and content information but keeps voice quality) yielded higher emotion recognition rates than those obtained through low pass filtering methods (which keep intonation contour intact but mask voice quality).

Support for the 'covariation view' that f_0 level, range and rate of change vary continuously with sympathetic arousal or activation is found in Pakosz (1983). This was confirmed in Bänziger and Scherer (2005) who found that the global descriptors of f_0 reflected primarily the level of arousal underlying portrayed emotions. Global movement of f_0 (global trend) was found to be significantly different for different emotions relative to the neutral state in Paeschke (2004) and was suggested as a suitable parameter to describe low activation emotions. The results reported in Paeschke, Kienast and Sendlmeier (1999) suggest that basic emotions (fear, anger, happiness, sadness, boredom, and disgust) are well differentiated in terms of global measures of f_0 mean and range, although differences in f_0 contour dynamics - the direction and steepness of f_0 changes (lowest in boredom and highest in anger) - were also reported. Ladd et al. (1985) reported that contour shape had

strong effect on the arousal scales: uptrend intonation (gradual rise, raised second peak) signals arousal, annoyance, and involvement.

Experimental evidence further suggests that f_0 alone (that is, without voice quality, intensity and timing cues) is not sufficient to communicate the entirety of affective message. For example, Lieberman and Michaels (1962) examined the contribution of f_0 in signalling emotion using acted expressions of boredom, confidentiality, disbelief, fear, happiness, pompous statement in content-neutral sentences and using forced choice paradigm. Resynthesised utterances in which only the smoothed f_0 and intensity envelopes were retained showed a drop in recognition rates from 85% to 47% relative to the natural utterances

Mozziconacci (1998) studied perceptual aspects of f_0 and speech rate variations[13] using natural (acted) speech, manipulations of natural speech through resynthesis and synthetic speech. Simple stylisation of f_0 by straight line approximation (voice quality and timing of the original utterance remained) did not impair good emotion identification. When the stimuli were generated based on the neutral utterance with the f_0 level, f_0 range, speech rate values for each emotion based on the production data, emotion identification, although still above chance level, deteriorated significantly. In particular, identification of sadness and fear was poor. This was largely due to the lack of emotion-specific voice source information. The author concludes that although pitch is important, the full emotional content cannot be conveyed only by pitch manipulation. Speech rate and voice quality must be taken into account.

The studies in Gobl, Bennett and Ní Chasaide (2002), Gobl and Ní Chasaide (2000, 2003b), Ní Chasaide and Gobl (2002, 2005) showed that synthetic stimuli incorporating voice quality manipulations, with or without f_0 variations, were perceived as affect-cueing by the participants in the listening tests. Conversely, stimuli in which voice quality was kept modal and only f_0 was varied depending on the affect, yielded much lower affective rating and were judged less expressive. The results support the findings in Mozziconacci (1998) and Scherer et al.(1984) as regards a somewhat weaker role of f_0 cues relative to voice quality. In Waaramaa et al. (2008) f_0 was eliminated by asking actors to express emotions (joy, tenderness, sadness, anger and a neutral state) on monopitched vowels [a:], [i:] and [u:]. Emotions were recognised with at least 50% accuracy showing that other prosodic expressive means (mainly related to spectral characteristics of the signal and measured in the study as alpha ratio and NAQ) successfully compensate for f_0.

[13] Source characteristics and intensity were not considered in her study.

On the other hand, the results of the listening tests reported in Carlson et al. (1992) suggest substantial contribution of f_0 to communication of emotional content. They used acted emotional speech (neutral, angry, happy, and sad) and systematically modified f_0 and duration through resynthesis. The manipulation involved different versions of time-alignment; for example, sad, angry and neutral utterances were all time-aligned to the happy utterance and resynthesised with the happy f_0 contour. The results of the forced choice listening tests demonstrated that the mixed stimuli, with the timing and f_0 taken from other sentences, were perceived as having the emotion of the sentence that contributed with the f_0 contour and segmental durations.

In summary, the research suggests little evidence of emotion-related f_0 contours; rather, f_0 varies continuously with the physiological arousal of the speaker. Voice quality and f_0 contribute independently to affect signalling and act independently of the verbal/linguistic content. Although f_0 variation is one of the most immediately discernible components in affect signalling, it is not sufficient for communicating emotional content in its entirety. Modelling f_0 variation based on the empirically obtained data of emotional speech does not yield high emotion recognition rating, mainly because voice source characteristics are not considered. When listeners hear the f_0 patterns associated with emotional speech, they do not very reliably identify the emotion. This production/perception mismatch is attributed to the lack of voice quality information. The suggestion in Frick (1985) that emotions are conveyed by 'prosodic contours' probably holds, when 'prosodic contours' are taken to include not only f_0 variation but all aspects of voice source dynamic variation.

This thesis aims at eliciting responses to stimuli that have strong f_0 variations with and without voice quality. In Gobl and Ní Chasaide (2003b) such f_0 variations were not included, which might account for the fact that strong emotions in particular were not signalled by the synthesised stimuli. The thesis aims at testing the assumption that if f_0 variation on its own is largely ineffective - it would yield more convincing results when combined with appropriate voice qualities.

CUE-CONFIGURATION AND COVARIATION OF ACOUSTIC CUES IN THE EXPRESSION OF EMOTIONS

The covariation (continuous) and cue-configuration (categorical) approaches discussed above apply to the whole field of affective acoustics: one approach suggests that acoustic associated with emotional states is best described by continuous dimensions of

valence and arousal; whereas the other claims that discrete emotion states correspond with a particular set of acoustic parameters.

For the cue-configuration approach, the key questions are whether different emotions are expressed through different combinations of acoustic cues and whether these acoustic parameters measure emotional states in a reliable way (Juslin & Scherer, 2005; Scherer, 2003). An apparent paradox seems to exist, however: whereas participants in listening tests are able to infer emotions from voice cues with better than chance accuracy, researchers have been unable to identify a set of cues that reliably differentiate among emotions (Juslin & Laukka, 2001). This could in part be explained by the limited use of relatively easily measureable acoustic parameters in emotion research. The use of a broader set of acoustic measures as well as more precise definitions of emotional states might show better emotion differentiation.

Another issue is the lack of data on the dynamics of the emotional speech signal: most studies give broad, averaged or global results. Fónagy (2001) argues that emotion differentiation based on the mean values of acoustic measurements is doomed to failure as the essentially dynamic nature of emotion expression is not adequately represented by such measures.

Despite considerable amount of empirical data, no definitive set of parameters has been established and the studies vary in the choice and interpretation of acoustic measures. For example, spectral tilt or jitter and shimmer can be measured in a number of ways. Certain parameters are more accessible methodologically than others, e.g., f_0, intensity and timing are relatively easy to measure, and have been used in the majority of studies. As appropriately observed by Juslin and Laukka (2001, p. 384), 'the amount of data available for a given cue is inversely related to the amount of difficulty involved in measuring and interpreting the cue'. However, with advances in digital speech signal processing, more sophisticated measures related to voice quality have become more widespread.

Recent studies of acoustic correlates of affective states appear to base their choice of parameters on the set used in an influential study by Banse and Scherer (1996), e.g., Aubergé and Cathiard (2003), Bänziger (2004), Bänziger and Scherer (2003), Drioli, Tisato, Cosi and Tesser (2003), Rossato et al. (2004).

In Banse and Scherer (1996), twelve actors were asked to portray 14 emotions varying in intensity and valence (hot anger, cold anger, panic, anxiety, despair, sadness, elation, happiness, interest, boredom, shame, pride, disgust, contempt) using two meaningless sentences composed from syllables of several languages. Listening tests showed better than chance accuracy of emotion recognition from these vocal expressions.

Acoustic analysis was conducted to obtain profiles of vocal parameters characterising the emotions: f_0, intensity, speech rate; LTAS measured for 9 bands of frequency (Hz): 125-200, 200-300, 300-500, 500-600, 600-800, 800-1000, 1000-1600, 1600-5000, 5000-8000; Hammarberg Index (the difference between the energy maximum in the 0-2000 Hz band and the 2000-5000 Hz band), and other measures. The results of acoustic analysis suggest the existence of differential emotion-specific vocal profiles for a large number of emotions: vocal parameters differentiate not only the intensity of emotions, but also valence or quality aspects. However, as was pointed out in Johnstone and Scherer (2000), the selection of parameters used in Banse and Scherer (1996) is not beyond critique and may require improvement: for example, it is not clear how the LTAS-based measures relate to vocal tract or glottal waveform characteristics.

Apart from Banse and Scherer (1996), evidence for the existence of emotion-specific vocal profiles has been reported, for example, in Johnstone, Banse and Scherer (1995), Scherer (2000a), Juslin and Laukka (2001). Summaries of the vocal cues characterising particular can be found in Banse and Scherer (1996), Frick (1985), Johnstone et al. (1995), Juslin and Laukka (2001, 2003), Laukka (2008), Murray and Arnott (1993), Scherer (1986, 1989, 2000a, 2003), Scherer et al. (2003), Klasmeyer and Sendlmeier (2000). Acoustic correlates of selected most commonly studied affective states based on the empirical research described in these studies are summarised in the Table 2.4 below.

The establishment of emotion-specific acoustic profiles is complicated by the fact that the same cue may be used in the same way to signal more than one emotion, e.g., mean f_0 is high in both anger and fear. Conversely, the same emotion may be expressed by different use of the same cue. For example, Aubergé and Cathiard (2003), in their study of the prosodic characteristics of real and acted amusement, analysed the stimuli using a set of parameters after Banse and Scherer (1996) and Mozziconacci (1998). The prosodic parameters found to be most telling in the expression of amusement are intensity and f_0 declination, but high inter-speaker variability in the expression strategies has been found. The authors conclude that there may be numerous ways of using the same parameters to express emotion.

Several authors have opted to measure the maximum of parameters with a subsequent application of statistical procedures, such as principal component analysis or automatic classification to identify the acoustic variables that carry most weight for emotion differentiation, e.g., Bänziger and Scherer (2003), Hammerschmidt and Jürgens

Table 2.4. *Vocal correlates of selected affective states (summary based on the literature).*

Affect	f_0	Intensity	Voice quality	Temporal and other characteristics
anger	high mean f_0; hot anger - high f_0 variability and range, cold anger - low f_0 variability and range, high proportion of rising f_0 contours	high mean intensity, increased intensity variability	'shouted modal' , tense, harsh voice, imperfect, breathy voice production, much high frequency energy, fast voice onset, 'steep' glottal waveform, much microstructural irregularity (jitter)	fast rate, low proportion of pauses, high precision of articulation, irregular rhythm
fear	high mean f_0 and f_0 variability, wide f_0 range, high proportion of rising f_0 contours	low mean intensity (although high in panic fear), high intensity variability	whispery, whispery or breathy falsetto, slow voice onset, 'rounded' glottal waveform, much high frequency energy, much microstructural irregularity (jitter)	fast rate
joy/ happiness	high mean f_0 and f_0 variability, wide f_0 range, high proportion of rising f_0 contours	medium to high mean intensity, high intensity variability	increase in high frequency energy, loud modal voice, 'breathy blaring' 'steep' glottal waveform, fast voice onsets, little microstructural irregularity (jitter)	fast rate, high precision of articulation
sadness	low mean f_0 and narrow f_0 range, little f_0 variability, mainly falling f_0 contours	low mean intensity	whispery, creaky, breathy voice, little high frequency energy, 'rounded' glottal waveform, slow voice onsets, microstructural irregularity (jitter)	slow rate, large proportion of pauses, low precision of articulation
boredom	low mean, floor and range of f_0, monotonous pitch	low mean intensity	creaky voice	slow rate, low precision of articulation

(2007), Monzo, Alías, Iriondo, Gonzalvo and Planet (2007), Schröder et al. (2001). For example, Hammerschmidt and Jürgens (2007) used an acoustic analysis system (LMA 9.2) that allows the extraction of 94 acoustic parameters, in an effort to provide more detailed description of prosodic characteristics of emotional speech. In their study, correlation analysis was used to reduce to 15 the initial 94 parameters extracted from repetitions of the word 'Anna' obtained from over twenty speakers expressing rage/hot anger, despair/lamentation, contempt/disgust, joyful surprise, voluptuous enjoyment/sensual satisfaction, and affection/tenderness. The results showed that each emotion can be characterised by a specific acoustic profile, differentiating among emotions. Specific LMA 9.2 parameters reported to differentiate between positive ('hedonistic') and negative ('aversive') emotions as groups were the ratio of peak frequency to fundamental frequency; the peak frequency; the percentage of time segments with non-harmonic structure; frequency range within single time segments, and time of the maximum of the peak frequency within the utterance.

Similar results are reported in Monzo et al. (2007) who focused on the capability of voice quality parameters to differentiate among a number of expressive speech styles (neutral, happy, sad, sensual and aggressive) using an expressive speech corpus in Spanish (female speaker). All parameters (local jitter and shimmer; HNR and Glottal-to-Noise Excitation ratio; spectral flatness measure (SFM); the Hammarberg Index; the drop-off of spectral energy above 1000 Hz; and the relative amount of energy in the high (above 1000 Hz) versus the low frequency range of the voice spectrum) were found important for the discrimination of expressive speech styles, with sensual and aggressive being the most clearly discriminated.

The above approach is preferred when the study focuses on the description of acoustic correlates of subtle and difficult to elicit affect-related phenomena (e.g., sarcasm). These could be difficult to quantify through the measures of f_0 and intensity, and careful consideration of the choice of acoustic parameters is particularly required. For example, Cheang and Pell (2008) attempted to identify possible acoustic cues of sarcasm based on recordings of English speakers producing sarcasm, humour, sincerity, and neutrality. Acoustic analyses included f_0 mean and standard deviation, f_0 range, mean amplitude, amplitude range, speech rate and HNR. Although the authors conclude that sarcasm was reliably characterised by overall reductions in mean f_0 and HNR relative to all other target attitudes, it is obvious that more sophisticated measures related to voice quality could have been useful.

The acoustic studies of vocal expression represent another case of a trade-off between the amount of data processed, the number of measures involved and the accuracy of the obtained results. For example, Bachorowski and Owren (1995) examined only three parameters - f_0, jitter and shimmer - in short speech samples (vowel segment from repetitions of the word 'test') obtained from a relatively large group of speakers (120) using an emotion induction technique. The participants received success or failure feedback while performing a lexical decision task and the results showed changes in all three measured parameters in response to induced positive (success) and negative (failure) emotional states. Experimental work that involves dealing with a large amount of data inevitably requires automatic processing and application of statistical measures in which a considerable part of the information on fine emotion-related variation of the voice signal may be lost. For example, Schröder et al. (2001)(Schröder et al., 2001) performed acoustic analysis of the Belfast Naturalistic Emotion Database semi-automatically using the ASSESS system which generates a simplified representation of the speech signal based mainly on the f_0 and intensity contours. A simplified description of the global trends of f_0 and intensity contours over time is done by identifying peaks and troughs in the contours and boundaries of pauses and fricatives; acoustic measures are then taken for segments between these landmarks (piecewise variables).

Experimental support for the covariation approach (that speech acoustics reflect continuous arousal and valence dimensions and not discrete states) - evidence for acoustic manifestation of emotion dimensions of activation, valence, and power – is reported in Bachorowski (1999), Bachorowski and Owren (2003), Murray and Arnott (1993), Schröder (2004a, 2004b), Schröder et al. (2001), Russell et al. (2003). To summarise their findings, activation is positively correlated with mean f_0, mean intensity, and, in most cases, with speech rate; as well as with f_0 range and high-frequency energy. The most reliable findings are related to arousal level. The evidence for valence and power is less pronounced. Positive valence corresponds to a faster speaking rate, less high frequency energy, low f_0 and wide f_0 range; a lax ('warm') voice quality; and longer vowel durations. An acoustic analysis of the Belfast Emotional Speech Corpus (Schröder et al., 2001) was conducted on the total of 26 parameters including the f_0 mean and range, the number of f_0 maxima/minima per time unit, the duration, magnitude and steepness of f_0 rises and falls; the duration of pauses, the articulation rate; the intensity mean and range; the spectral slope and the approximations of the Hammarberg indices. Results showed that nearly all acoustic variables show substantial correlations with the emotion dimensions, and in particular with activation. Laukkanen et al. (2008) report some evidence of the importance of F2 and F3 in

the perception of valence. In particular, vowel samples with F3 artificially raised by 30 per cent were perceived more often as expressing positive affect than the samples of the same vowels with the original, artificially lowered or removed F3.

Bachorowski and Owren (2008) state categorically that neither the cue-configuration nor the covariation approach is strongly supported by empirical evidence. They suggest a perspective entirely different from the models above. According to their 'affect-induction model', the fine variations of speaker emotional states are not expected to be correlated with particular vocal cues; the speaker/vocaliser actively uses acoustic features not to express emotion per se, but to manipulate listener's arousal, affective state and behaviour in a way beneficial for the vocaliser [14]. Bachorowski and Owren (2003, 2008) maintain that 'rather than looking for the correlations between emotional states and acoustic cues, research needs to focus on the effects that sounds have on listeners' emotional state.

[14] Similar is suggested in Russell et al. (2003), Zinken, Knoll and Panksepp (2008), Kappas and Poliakova (2008).

2.4 Universality and culture specificity of vocal expression of emotion

The non-verbal expression and communication of emotion is shaped by cultural factors (for reviews, see Izard, 1980; Lehman, Chiu, & Schaller, 2004; Mesquita, Frijda, & Scherer, 1997; Oatley et al., 2006; Ogarkova et al., 2009; Porter & Samovar, 1998; Scherer & Brosch, 2009; Vinden & Astington, 2000). These cultural factors may include everything that groups of people have in common: language, race, political systems, geography, beliefs and practices, behavioural patterns, value systems, language etc, in fact all except very basic biological needs and reflexes (Scherer & Brosch, 2009).

In general, the culture-specific influence on emotion expression/communication is opposed to universal biological factors (Zinken et al., 2008). On the one hand, vocal expression of biologically based ('hardwired', 'basic') emotions represents a universal (and often involuntary) behaviour, although in certain cases modified by display rules. Vocal expressions of 'basic' emotions related to distinct physiological changes such as muscle tension and sympathetic arousal (fight-or-flight response) tend to be universal, that is expressed similarly across different cultures and universally recognised. Thus, the biological component is universal. On the other hand, there is the vocal communication/expression of more complex cognitive emotions (affective states), often for manipulative purposes, when affect communicated is not always affect experienced (e.g., when parents use 'angry' tone-of voice to discipline their children without necessarily experiencing anger as such or when people use 'civilised' neutral tone-of-voice while internally experiencing, but masking or suppressing anger). Here cultural factors govern the 'display rules' and conventions of emotional expression (be it true or fake). Evidence of culture specific appraisal of the same eliciting events has also been reported (e.g., Oatley et al., 2006; Scherer & Brosch, 2009). Thus, the cognitive component is culturally informed.

Different cultures may vary in the way they ritualise the display of emotions (different stereotypical iconic expressions of the same emotion may exist in different cultures) as well as in ways they regulate expressive behaviour according to culture-specific display rules. Cultures differ in the general attitude towards or relative importance they attach to specific emotions. They also differ in the extent to which they 'allow' emotions to be expressed, vocally or facially. Within a given culture, some emotions are seen as admirable, others may be despised. An emotion may be 'hypercognised' (strongly emphasised and valued) by a particular culture, which is manifested in a particularly rich

vocabulary describing this emotion. Members of this culture are considered more likely to experience and to differentiate fine shadings of this emotion (Levy, 1984; Mesquita, 2003; Mesquita & Ellsworth, 2001; Oatley et al., 2006; Prinz, 2004b). The above concerns all aspects of the nonverbal expression and communication of emotion, including the vocal expression.

Cultures vary considerably in terms of the relative importance they attach to the non-verbal communication of emotion compared to the verbal content of the message. According to Kitayama and Ishii (2002), who follow the description first proposed by Edward Hall (1976), this difference is especially prominent between individualist languages and cultures and collectivist ones. For example, in 'low-context' Western cultures and languages (e.g., English or German) they claim that a greater proportion of information is conveyed by verbal content, while nonverbal cues such as tone-of-voice play a relatively minor role. They claim the opposite to be true for certain East Asian and Mediterranean cultures and languages (e.g., Japanese, Greek or Spanish), that are considered 'high context', where nonverbal cues are proposed to be of greater importance relative to the verbal message. Thus individuals belonging to high-context cultures are claimed to develop an 'attentional bias' to non-verbal vocal cues and appear to be more sensitive to subtle tone-of-voice variations, which can be illustrated, for example, by the use of creaky voice in Japanese (Kitayama & Ishii, 2002).

The cross-cultural perception of emotion from facial expression has been comprehensively studied (Ekman, 1993, 1982; Ekman et al., 1987; Elfenbein & Ambady, 2002b; Hager & Ekman, 1983). Facial expressions commonly associated with the basic emotions (such as anger, disgust, fear etc.) have been shown to be expressed and recognised similarly by members of different cultures. In contrast, the problem of universality of emotion expression in voice remains unresolved, mainly due to the lack of systematic and representative studies across different cultures. Obtaining comparable emotional speech samples from representatives of different languages/cultures is difficult, further difficulties are related to ethical considerations as well as the problem of translation equivalence of emotion-denoting words (e.g., Matsumoto & Assar, 1992; Mesquita & Frijda, 1992; Russell, 1991).

Cross-cultural research of non-verbal behaviour other than facial expression has been attempted through the use of questionnaires and self-reports, e.g., Scherer and Wallbott (1994), Walbott, Ricci-Bitti and Bänninger-Huber (1986). Walbott et al. (1986) conducted a questionnaire-based cross-cultural study of emotional experience and reaction of joy, sadness, fear, and anger conducted in eight European countries (Belgium, France,

Great Britain, Israel, Spain, Switzerland and West Germany). Negative emotions were found to be generally more easily recognised from voice cues than positive emotions, with anger being best recognised. Joy was found to be most expressed in non-verbal behaviour, followed by anger and sadness, with fear being the emotion most expressed in physiological symptoms. The non-verbal reactions were reported to be controlled more when negative emotions were experienced. These patterns were found independent of country differences, providing some evidence for the universality of non-verbal expression. The authors conclude that there seem to be general display rules that imply that positive emotions have to be controlled to a lesser degree than negative emotions.

Cross-cultural perception studies suggest broad cross-cultural similarities in affect perception from voice, e.g., McCluskey and Albas (1981), McCluskey, Albas, Niemi, Cuevas and Ferrer (1975), Scherer et al. (2001), van Bezooijen, Otto and Heenan (1983), Graham, Hamblin and Feldstein (2001). However, a clear effect of 'in-group advantage' has been frequently reported, that is, the degree of accuracy of emotion recognition/attribution is higher when listeners judge emotion expressions produced by the speakers of their own language. Elfenbein and Ambady (2002b) conducted a meta analysis of universality and cultural specificity of emotion recognition within and across cultures based on 97 studies (42 different nations, 23 different ethnic groups). They report evidence for an in-group advantage across a range of experimental methods, positive and negative emotions, and different nonverbal channels of communicating emotion, such as facial expressions, tone of voice, and body language.

The in-group advantage and poorer cross-language recognition of expressive cues from voice could possibly be explained by the general neutral tone-of-voice of an unfamiliar language or the effects of a certain linguistic features unrelated to emotional expression such as the linguistic use of a particular phonation type which could be taken as an affective vocal cue for the listeners from a different culture. The in-group advantage is further contributed to by the absence of distortions in the course of translation of emotion-denoting words (Elfenbein & Ambady, 2002a, 2002b; Mesquita & Frijda, 1992).

Elfenbein and Ambady (Elfenbein & Ambady, 2002b, 2003) further report varying degree of cross-cultural accuracy in emotion recognition depending on the particular emotion tested as well as on the modality of emotion expression. Happiness was most accurately recognised from the face but least accurately from the voice. Anger and sadness were most accurately recognised cross-culturally from the voice, but relatively less accurately from the face. The in-group advantage was smallest for happiness in the face, but the highest for happiness in the voice. Fear showed a relatively large in-group

advantage in both the face and the voice, whereas anger showed a relatively small in-group advantage in both modalities. The cross-cultural accuracy of emotion recognition is reported to vary depending on experimental design (was found to be lower in studies that used a balanced research design) and the types of the stimuli used (the accuracy is higher in studies that used imitation rather than spontaneous emotional expressions). The authors conclude that empirical evidence points at the existence of 'universal and biologically based core components of emotions'; however, there is also evidence that emotion communication is shaped by cultural factors.

Scherer et al. (2001) points out that empirical research on cross-cultural emotion recognition from the voice is extremely rare. Earlier studies are reported in Albas, McCluskey and Albas (1976), Beier and Zautra (1972), Kramer (1963, 1964), McCluskey and Albas (1981), McCluskey et al. (1975), van Bezooijen et al. (1983). More recently the problem was discussed in Abelin (2004), Abelin and Allwood (2000), Biadsy, Rosenberg, Carlson, Hirshberg and Strangert (2008), Breitenstein et al. (2001), Burkhardt et al. (2006), Campbell (2004b), Erickson (2005), Linnankoski et al. (2005), Mathon, de Abreu and Perekopska (2006), Sawamura et al. (2007), Scherer (2000a), Scherer et al. (2001), Schröder, Heylen and Poggi (2006), Shochi, Aubergé and Rilliard (2007), Teshigawara, Amir, Amir, Milano Wlosko and Avivi (2009). These will be briefly summarised below.

A number of studies by Balkwill and colleagues describe cross-cultural recognition of emotive improvisations expressing joy, sadness, anger in music (Balkwill, Cuddy, & Thompson, 2004; Balkwill & Thompson, 1999; Balkwill, Thompson, & Matsunaga, 2004). Their results showed significant main effects of intended emotion and musical culture on emotion attribution. Despite broad similarities, the relative influence of specific acoustic cues (mean and range of intensity; mean and range of f_0, complexity/event density) differed for listeners belonging to different cultures.

The studies concerned with cross-cultural perception of emotion generally follow the standard paradigm of emotion decoding studies using acted portrayals of a limited set of discrete emotion categories on content free utterances and report the accuracy of perception of emotions expressed by speakers of one culture by the listeners from another culture. Similarity in emotion attribution across cultures has been shown, for example, by Scherer et al. (Scherer, 2000a; Scherer et al., 2001). In their large scale study, vocal emotion portrayals of anger, sadness, fear, joy and neutral voice by professional German actors in content-free specially constructed sentences were judged by listeners from Germany, Switzerland, Great Britain, Netherlands, US, Italy, France, Spain, and Indonesia. The results suggest that judges from different countries made similar assumptions about

emotions and made the same mistakes in emotion attribution (an overall accuracy of 66 per cent across all emotions and countries and similar patterns of confusion for all countries were reported). Evidence on similarities in cross-language judgement of emotion expression is reported also in Campbell (2004b), Sawamura et al. (2007).

The empirical findings point at different ways in which the speakers of different languages express and recognise vocal emotions: the same emotion may be expressed 'more clearly' and recognised with higher accuracy by representatives of a certain culture compared to others. For example, in the studies by McCluskey et al. (McCluskey & Albas, 1981; McCluskey et al., 1975) the Mexican subjects were found to be significantly more sensitive to expressions of happiness, sadness, love, and anger in speech than their Canadian counterparts. Emotions expressed by Mexicans were generally recognised with higher degree of accuracy by both Mexicans and Canadians than those expressed by Canadians. In a study by Albas et al. (1976) following the same paradigm, electronically filtered emotion expressions by white and Indian adult males were presented to groups of white and Indian listeners. The white listeners were able to identify the emotional content of tone of voice more accurately than the Indian listeners; the Indian listeners performed more accurately when judging the samples from Indian speakers (in-group advantage).

Similar language- and culture-related bias in the listeners' responses is reported in Abelin and Allwood (2000) where an utterance by a Swedish speaker expressing joy, surprise, sadness, fear, shyness, anger, dominance, disgust was judged by Swedish, English, Finnish and Spanish listeners in a free-labelling task. Emotion recognition ratings were high in the case of the Swedish, English and Finnish listeners, but attribution was a lot less accurate in the case of the Spanish. For example, 35 per cent of the Spanish speakers interpreted Swedish expression of happiness as sadness and 30 per cent of the Spanish speakers interpreted Swedish expression of surprise as anger. Sadness and anger, however, were interpreted mostly as targeted. Spanish and Finnish listeners interpreted Swedish fear as sadness. In Abelin (2004), acted emotional expressions (sad and tired, angry, sad, sceptical, delighted, afraid, depressed, very happy) uttered by a Spanish speaker were judged by Swedish and Spanish speakers. The results, analysed in comparison with Abelin and Allwood (2000), showed that sadness is interpreted best cross-linguistically; Spaniards recognise Swedish anger well, but not the other way around. Happiness and fear were poorly recognised cross-linguistically by both groups, whereas sadness is best recognised from vocal expression. In general, Spanish recognise Swedes more accurately than the other way around. Spaniards poorly recognise Spanish anger/sadness compared to Swedish anger/sadness. The results for the Swedish listeners clearly show in-group bias.

Mathon et al. (2006) reported culture-specific differences in the recognition rates of expressions of anger extracted from a corpus of spontaneous speech in French and subsequently masked with white noise. Recognition rates of the French and Czech subjects were found to be higher than those of the Portuguese participants. French speakers appeared to be more sensitive to weaker degrees of anger which speakers of Czech and Portuguese recognised as 'no anger' (another case of in-group bias).

Language-based differences accounted for variability in emotion attribution reported in a large scale study by Scherer et al. (Scherer, 2000a; Scherer et al., 2001)[15]. The dissimilarity in emotion identification increased with increasing language dissimilarity from German. Speakers of Germanic languages (Dutch and English) showed the highest accuracy rates, followed by the speakers of Romance languages (Italian, Spanish, French), and with Indonesian speakers showing the lowest.

Schröder et al. (2006) compared the perception of affect bursts (for 10 emotions) uttered by a German speaker by German and Dutch subjects. While Dutch and German listeners yielded rather similar perception ratings, the authors reported a number of clear language-specific differences, e.g., the German expressions of admiration and disgust are not differentiated by the Dutch listeners.

The accuracy of cross-cultural emotion recognition has been reported to be age dependent (e.g., McCluskey & Albas, 1981; McCluskey et al., 1975) as well as gender dependent, e.g., Hatta and Nachshon (1988). In Hatta and Nachshon (1988), Japanese and Israeli subjects rated the intensity of the positive and negative emotional connotations of synthetically constructed nonsense sentences delivered to one ear while hearing white noise in the other ear. The results indicated that female listeners evaluated negative overtones as more intense, whereas for male listeners positive and negative overtones were rated as equally intense.

Cross-cultural studies also report emotion-specific degree of decoding accuracy. Kramer (1964)[16] investigated methods of judging the emotional load of content free utterances, such as speech in a language unknown to the listener. Portrayals by Japanese speakers of anger, contempt, grief, indifference, and love were presented to American listeners in a forced choice paradigm. The Japanese actors reported that grief was the easiest emotion for Japanese to express openly, followed by anger and contempt, and love being most difficult. The result of the listening tests showed recognition accuracy well

[15] The listeners were judging simulated emotions produced by German actors in artificially constructed sentences.

[16] He also presents a comprehensive literature review on methods of studies of perception of emotions from voice to date in Kramer (1963).

beyond chance, with grief, indifference and anger showing the highest degree of attribution accuracy, and contempt least often judged correctly. Love was often mistaken for indifference.

The accuracy of emotion attribution was found affect-dependent and culture-dependent in Beier and Zautra (1972). In their study, vocal expressions of the affect-related states (happy, fear, sad, anger, indifference, and flirt) varying in duration recorded by American speakers were presented to American, Polish and Japanese listeners. The best recognised for the Polish subjects were angry and indifference; and for the Japanese subjects, sad and indifference; for the American subjects, happy and anger. Non-American subjects had most difficulty in decoding the affects flirt and happy.

The results reported in Van Bezooijen et al. (1983) also suggest language-dependent reactions to the same emotional expression. In their study, emotions expressed by Dutch speakers were recognised differently by listener judges from Denmark, Taiwan, and Japan. For example, for the Taiwanese listeners, the best recognised were sadness and surprise; the Japanese subjects showed the highest recognition rate of sadness and the least recognition of shame. For the Dutch subjects, the expression best recognised was happiness, and the vocal expression of contempt showed the lowest recognition accuracy.

Cross-cultural studies present evidence on different sensitivity to various acoustic cues ('cue utilisation') by speakers of different countries. Burkhardt et al. (2006) studied the effects of varying pitch range, duration, and jitter simulation in MBROLA synthesised stimuli on the perception of emotionally loaded sentences (neutral, joyful, friendly, threatening, bored, frightened, sad) by the speakers of French, German, Greek and Turkish. The frightened utterance was found to be associated with narrow pitch range for French and Greek listeners, but not for German and Turkish ones. In contrast to the speakers from other countries, the Turkish associated friendly expression with a narrow pitch range. German and French listeners did not associate increased amount of jitter with the threatening affect, unlike the Turkish listeners. For the French listeners the impression of frightened was enhanced by jitter-simulation, whereas for the Germans the tendency was the opposite. The jitter-simulation generally lowered the acceptability of the stimuli for the Greek and Turkish listeners. Although not all effects can necessarily be explained by language- or culture differences alone, the authors conclude that emotion simulation in synthesis in global terms without culture-specific adjustments will not be adequate[17].

[17] In a similar vein, (Schröder, 2009) points out that the assumption of universality of global prosodic parameters often taken in expressive synthesis holds only as long as the number of available emotion categories is small.

Breitenstein et al. (2001) studied the contribution of speech rate and f_0 variation to emotion perception by German and American listeners. The listeners were presented with sentences in German uttered with five different emotions as well as the versions of the original emotional stimuli in which speech rate or f_0 were systematically altered. The results were generally similar for the two groups of listeners, but certain differences were also observed. For example, the happy stimulus in which speech rate was increased was associated with happy affect by Germans but the American subjects associated it with angry or frightened. The American subjects appeared to be more affected by rate manipulation than the German ones.

Biadsy et al. (2008) found similarities as well as culture-specific correlations between acoustic variables (f_0 mean and standard deviation, f_0 range, mean intensity and duration) and the way speech samples were rated in terms of charisma in standard American political speech by Americans, Palestinians and Swedes. Another experiment involved Americans and Palestinians judging Palestinian Arabic speech. For example, American and Palestinian subjects showed different preferences regarding speaking rate: a faster speaking rate in Arabic speech tokens was associated with charisma by the American raters, but not by the Palestinian ones.

Cross-cultural expression Johnstone and Scherer (2000) point at the lack of empirical evidence as regards the strategies members of different cultures employ in expressing the same emotion. Although a comprehensive study of this kind has yet to be conducted, observations by various authors suggest that production of expressive speech is undeniably affected by the language and culture background of the speaker. Auditory analysis based descriptions of voice quality show that the expressions of the same emotion vary across languages/cultures. For example, according to Burkhardt and Sendlmeier (2000) and Schröder (2001), who based their analysis on the data from German speakers, fear is associated with falsetto voice, whereas the auditory analysis of the speech of an English actor reported in Boula de Mareüil, Célérier and Toen (2002) showed that the actor used whispery voice to portray fear. Drioli et al. (2003) report that expressions of fear in Italian were associated with breathy voice. Creaky voice is associated with boredom in English and German (Burkhardt & Sendlmeier, 2000; Laver, 1980). However, in Japanese, the use of pressed or creaky voice ('*rikimi*') follows very complex rules of social interaction, and creaky voice has been reported to convey a range of affective states, e.g., surprise, irritation, disgust, anger and admiration (Sadanobu, 2004). The study of acoustic manifestations of anger, fear, joy and sadness in the Chinese emotional speech database reported in Yuan, Shen and Chen (2002) showed that anger is expressed with creaky voice.

Drioli et al. (2003) studied acted emotional speech in Italian and reported that a creaky voice quality was found in expressions of disgust. Collecting speech samples of acted emotion expression in English, French and Spanish, Boula de Mareüil et al. (2002) made interesting observations regarding language-specific strategies in the use of voice quality and pitch that the actors adopted in emotion portrayals. For example, when portraying fear, the English actor used voice quality (whispery voice), whereas the French and the Spanish actors rather shouted and cried respectively. The Spanish actor with normally (emotionally neutral) low-pitched voice was reported to dramatically modify his timbre and mean pitch to communicate the various emotions. While these could be merely idiosyncratic strategies, they are possibly also indicative of language-specific stereotypes of vocal emotion expression. Further evidence of culture-specific voice stereotypes is reported in (Behlau & Gasparini, 2009; Teshigawara et al., 2009)

A study of f_0 range, peak alignment and speech rate in 4 monolingual (French) and 4 bilingual (French/English) children (Grichkovtsova & Mennen, 2008) suggests that bilingual children use these parameters differently in emotion expressions (e.g., happiness) in their two languages.

In summary, vocal expression/communication of emotion is shaped by both biologically based and culturally informed components. While biologically based 'simple' emotions are expected to be expressed universally across cultures, expression of more complex mental states and emotions with substantial cognitive component is modified by culture-specific display rules. Empirical evidence collected to date suggests that emotional information expressed by speakers of one language/culture can be recognised from voice alone by speakers of other languages. This implies the existence of universally recognisable characteristics of vocal patterns of emotion. No comprehensive study into cross-cultural encoding of emotion has been conducted to date, although impressionistic observations point at notable language-specific differences. These differences need to be studied in a systematic manner.

2.5 Summary

The aim of the chapter was to provide theoretical background underlying the study of voice and emotion as well as summarise some important empirical findings reported to date. It further aimed at providing theoretical background for a series of experimental studies reported in Chapters 3-6 of this thesis.

The problems related to defining and measuring voice quality as well as the role of voice in emotion expression was discussed. Voice quality can be defined as a quasi-permanent characteristics colouring of a speaker's tone of voice. Although defined as an essentially perceptual phenomenon, it has rarely been studied using psychoacoustic methods. Auditory analysis of voice quality can be conducted using a set of labels describing discrete types of voice quality. However, this approach implies the existence of fixed internal standards shared by all trained observers and usually results in low degree of interrater agreement. A number of parameters based on the acoustic analysis of speech waveform and glottal volume velocity waveform that are commonly applied in objective measurement of voice quality have been outlined. No single parameter is sufficient to capture all characteristics of voice quality; therefore, a combination of acoustic parameters is usually required for voice quality measurement.

Voice is an important carrier of linguistic, paralinguistic and extralinguistic information. Paralinguistically, it is used to signal the information on the affective state of the speaker, including emotion, mood, and attitude. Encoding of emotional information in the voice is shaped by both 'push' and 'pull' effects, that is by emotion-related physiological processes and external factors related to socially accepted norms.

Vocal expression of emotion has been studied extensively, and a number of theoretical approaches have been adopted by researchers in the field, most commonly the basic emotion theory and the dimensional theory. Component process model within the appraisal theoretical framework has also been applied. Traditionally, the emotion communication process has been viewed as involving the stages of emotion encoding, transmission of emotionally coloured acoustic signal and emotion decoding. However, in recent years the focus tends to be shifted from this rather simplistic representation towards treating expression of emotions as a means by which the speaker can manipulate the affective state of the listener.

The studies into vocal expression and communication of emotion usually involve emotional speech data collection (naturalistic databases, elicited or acted emotions) which is subsequently presented to listener judges to evaluate their ability to accurately recognise

(decode) expressed emotions, in a forced choice or free-labelling task. Perceptual experiments are also used to evaluate the relative contribution of acoustic cues, such as f_0, intensity, speech timing characteristics and voice quality, to the vocal expression of emotion. In such experiments, certain acoustic cues could be masked or manipulated through resynthesis.

The studies of acoustics of emotionally coloured speech have been focused on establishing emotion-specific combinations of acoustic parameters (emotion-specific 'acoustic profiles') as well as acoustic correlations of such emotion dimensions as activation and valence. Due to methodological difficulties, the characteristics of the voice source in vocal expression of emotion is considerably less studied compared to traditional and conceptually less problematic measures related to f_0, intensity and timing. The results of the studies of acoustic properties of emotional speech can inform the endeavours to generate emotional speech synthesis. It is increasingly obvious that natural sounding speech synthesis will require being able to replicate the kinds of affect related tone of voice variations that are used in human speech communication.

In general, the culture-specific influence on emotion expression/communication is opposed to universal biological factors. The expression of hard-wired biologically based emotions is considered universal, whereas the vocal communication/expression of more complex cognitive emotions is governed by culture-specific 'display rules' and conventions.

Large scale cross-cultural studies are extremely rare, primarily due to numerous restrictions in terms of ethics, feasibility, and cost. Although no systematic study into the way emotions are encoded across languages, auditory analysis based studies show clear language-specific strategies that the speakers adopt in emotion encoding. The cross-cultural studies of emotion recognition from voice usually follow the established paradigm of decoding studies, and deal mostly with the ability of the listener judges of one country (countries) to recognise simulated emotion speech samples uttered by representatives of other language/culture. The empirical evidence points at both similarities and culture-specific features in emotion attribution across cultures. The same emotion may be expressed 'more clearly' and recognised with higher accuracy by representatives of a certain culture compared to others; and the speakers of different languages show different degree of sensitivity to the same acoustic cue.

3 Chapter three
Cross-language study: methodology

3.1 Introduction

Despite the growing body of research dedicated to cross-cultural aspects of vocal expression and communication of emotion (as discussed in Chapter 2) much more research is needed to establish to what extent the vocal cues to affect are common to particular languages, and to what extent they are not. One of the main objectives of this thesis is to explore cross-language/cultural differences and similarities in how tone of voice maps to affect. A series of early perception studies conducted at the Phonetics and Speech Laboratory, Trinity College Dublin have demonstrated that different voice qualities are robustly associated with particular affective overtones, for Hiberno-English listeners. See, for example, Gobl, Bennett and Ní Chasaide (2002), Ní Chasaide and Gobl (2002, 2005), Gobl and Ní Chasaide (2000, 2003b), Gobl (2003).

The principal question addressed in this study, and the subject of Chapter 4, concerns the extent to which these associations hold for listeners from different language/cultural backgrounds. This was explored through a series of perception experiments on groups of Spanish, Japanese, Russian and Hiberno-English listeners, using essentially the same type of stimuli as were used in those earlier studies mentioned above, and using a similar research paradigm. The types of the stimuli and the steps involved in their generation are outlined below in Section 3.2.

Prior to the description of the stimuli and research paradigm, certain preliminary comments need to be made that motivate the choice of the stimuli.

The main objective of the study is to explore the mapping of voce quality to affect in each of the four language groups. The listeners' reactions were elicited not to individual dimensions of voice quality, such as spectral slope or even breathiness, but rather to the holistic voice quality entity, such as breathy voice or tense voice. Therefore, the first goal with the stimuli was to produce an utterance that could be rendered with a range of distinct voice qualities. This involved complex manipulations to the stimulus utterance in ways that would render holistic impressions of particular voice qualities. Individual voice qualities involve a number of source features working together (see Laver (1980), Gobl (1988, 1989), Gobl and Ní Chasaide (1992) and the general review in Chapter 2). The particular manipulations relevant to this study are detailed below in Section 3.2.

The synthesis was guided by the earlier analytic studies just mentioned based on recordings of a phonetician trained in the production of voice qualities of the Laver's

framework (Laver, 1980). It was also partially guided by the broader literature on voice quality. The synthesis was of a short utterance in Swedish, semantically neutral for the linguistic groups of this study. The source and filter manipulations are carried out in a way that ensured that the voice source dynamics were in keeping with that found in natural utterances (see Figure 3.1 below). As pointed out in Gobl and Ní Chasaide (2002, 2003b) and in Fónagy (2001), the dynamics are particularly crucial.

The set of voice qualities chosen for synthesis included modal voice, whispery voice, breathy voice, lax-creaky voice, tense voice (and harsh voice, for the experiment in Chapter 5). Voice qualities are not discrete entities, but may vary in a continuum, just as vowel qualities vary continuously in the vowel space. Therefore, it was important to work within a well defined reference system, just as segmental phoneticians work with the Cardinal Vowel system. The Laver's taxonomy (Laver, 1980) offers such a reference framework.

The set of qualities used was of necessity a limited sampling of a potentially very large variety of possible voice qualities. Although the set chosen was intended to give good coverage of the main voice qualities which are most frequently referred to in the literature, there are necessarily omissions (e.g., falsetto). The voice qualities were also not intended to be extreme exemplars of their type. The synthesised voice quality stimuli were informally evaluated by a phonetician trained in the Laver system as being representative of their type.

Motivating f_0 stimuli, on their own and in combination

Much research has been dedicated to the role of f_0, and it is well established that f_0 cues are important for affect signalling (see Chapter 2, Section 2.3.3). In many works, Scherer and colleagues (Banse & Scherer, 1996; Grandjean et al., 2006; Ladd et al., 1985; Scherer et al., 1984) have detailed how the f_0 level, range and dynamics are correlated with affect in speech production. However, the same f_0 features, when integrated for perception studies, appear not to cue affect very reliably. Replicating production values of f_0 alone in prior research (Bänziger & Scherer, 2005; Mozziconacci, 1998 and others discussed in Chapter 2) has not been found to effectively cue affect. Most researchers conclude that voice quality is likely to be the missing link: this apparent mismatch of production and perception might be explained by the fact that voice quality cues were missing in such perception experiments in which only f_0 is modelled.

In real speech the voice source varies as a whole in the expression of emotion, and large f_0 perturbations are likely to occur with concomitant shifts in voice quality. Earlier

perception studies using stimuli similar to those employed here (Gobl & Ní Chasaide, 2003b) showed that these voice qualities were better at signalling milder affects than strong emotions, with the exception of anger. It was hypothesised that the signalling of strong emotions would require concomitant large f_0 perturbations, such as those described in the literature (Banse & Scherer, 1996; Mozziconacci, 1998).

This hypothesis was tested in Gobl et al. (2002) by including a series of stimuli where along with a basic set of voice quality variants, a combined set was tested, where the same qualities incorporated f_0 contours that were likely to be appropriate. Results largely supported the hypothesis in that the stronger emotions did appear to be more strongly associated with the combined stimuli. A difficulty with this study is that, when a combined stimulus was effective, it was not clear whether it was as a consequence of f_0 differences per se, or of the combining of f_0 with voice quality information.

The present study further explores these issues. First, it was of interest to ascertain whether voice quality cues (particularly to strong emotions) become more effective when major f_0 perturbations are included. Secondly, I wished to evaluate the contribution of the same f_0 contours on their own, without voice quality variations. Finally, it was of interest to test whether and to what extent the different language groups might differ in their handling of these different possibilities.

For that reason, in addition to the stimuli varying in voice quality, two further series were generated. The first of these, 'f_0 only' group simply involved manipulations to the original modal voice stimulus, so as to incorporate a variety of f_0 contours. A further group of stimuli combined specific f_0 contours and voice qualities most likely to co-occur. The contours chosen were based on Mozziconacci (1995) and are described in Section 3.2.

A detailed description of the research methodology and the stimuli used in the experiment is given below.

3.2 Materials: the synthesised stimuli

The stimuli used in the present cross-language study are, with minor modifications, the ones described in detail in Gobl and Ní Chasaide (2003b). The stimuli were based on a recording of a short Swedish utterance 'ja adjö' ['jɑ a'jø] produced with modal voice by a male speaker. This was analysed and resynthesised (see below). For the synthesis, the LF model implementation of the KLSYN88a formant synthesiser (Sensimetrics Corporation, Boston, MA, described in Klatt and Klatt (1990)) was used. From this first modal version of the utterance three types of stimuli were generated.

(1) 'VQ only' – stimuli with different non-modal voice quality settings but in all of which the original, neutral f_0 contour was used. The different voice qualities were produced by manipulating voice source parameters and (where appropriate to the voice quality) filter parameters.

(2) 'f_0 only' – stimuli in which modal voice was combined with different affect-related f_0 contours. The f_0 contours were adapted from (Mozziconacci, 1995).

(3) 'VQ + f_0' – stimuli in which the voice qualities of (1) were combined with the specific affect-related f_0 contours from (2).

3.2.1 'VQ only' stimuli: modal voice

The non-modal voice stimuli of the 'VQ only' series were synthesised on the basis of a high quality copy synthesis of the utterance 'ja adjö' ['jɑ a'jø] produced by a male speaker with the modal voice quality as described in Laver (1980). Pulse by pulse inverse filtering of the utterance (106 pulses) was done using the manual interactive technique described in Gobl and Ní Chasaide (1999a), Ní Chasaide, Gobl and Monahan (1992). Subsequent parameterisation was done by fitting the LF source model (Fant et al., 1985) to the estimated differentiated glottal flow using the same interactive software (Gobl & Ní Chasaide, 1999a). This parameterisation produces data for the source parameters EE, RA, RG and RK, described in Section 2.1.3 of Chapter 2, Table 2.1.

A copy synthesis of the utterance was generated with the LF model implementation of KLSYN88a formant synthesiser. Note that this synthesiser incorporates the LF model (in a slightly modified version). In KLSYN88a, the control parameters AV, TL, OQ, SQ, although called differently, are direct transforms of the LF-derived parameters and represent essentially the same information. The mapping from LF-derived parameters to

KLSYN88a parameters is presented in Table 3.1 and these are fully discussed in Mahshie and Gobl (1999), Gobl (1988), Klatt and Klatt (1990).

In the copy synthesis of the modal utterance, 14 synthesis parameters were varied dynamically: frequencies of the first five formants F1, F2, F3, F4, F5; bandwidths of the first and the second formants B1 and B2; fundamental frequency F0, amplitude of voicing AV, spectral tilt TL, open quotient OQ, speed quotient SQ, aspiration noise AH, and diplophonia DI. For practical reasons, the input data was reduced by selecting values at specific time points for each parameter, the number of values ranging between 7 and 15, depending on the parameter. Linear interpolation was used between the selected points to capture the natural dynamics of the signal[18]. The utterance resulting from the copy synthesis of the original recording was used as a modal voice stimulus (Gobl & Ní Chasaide, 2003b; Ní Chasaide & Gobl, 2005).

Table 3.1. *LF to KLSYN88a parameter transformation.*

Parameter	**LF model**	**KLSYN88a**	**LF model to KLSYN88a parameter mapping**
Amplitude of voicing	EE	AV	$AV = 20 \cdot \log(EE)$
Spectral tilt	RA and f_0	TL	$TL = 10 \cdot \log\left(1 + \left(\frac{3000}{FA}\right)^2\right)$ $where\ FA = \frac{1}{2 \cdot \pi \cdot TA}$
Open quotient	RG and RK	OQ	$OQ = \frac{(1+RK)}{(2 \cdot RG)}$
Speed quotient (glottal skew)	RK	SQ	$SQ = \frac{1}{RK}$
Aspiration noise	-	AH	
Diplophonia (double pulsing)	-	DI	

3.2.2 'VQ only' stimuli: non-modal voice qualities

Five non-modal voice stimuli – whispery, breathy, lax-creaky, harsh and tense - were generated on the basis of the modal voice by manipulating the seven source parameters (fundamental frequency F0, amplitude of voicing AV, spectral tilt TL, open

[18] These complex time-varying dynamics are shown in Figure 3.1 as parameter trajectories.

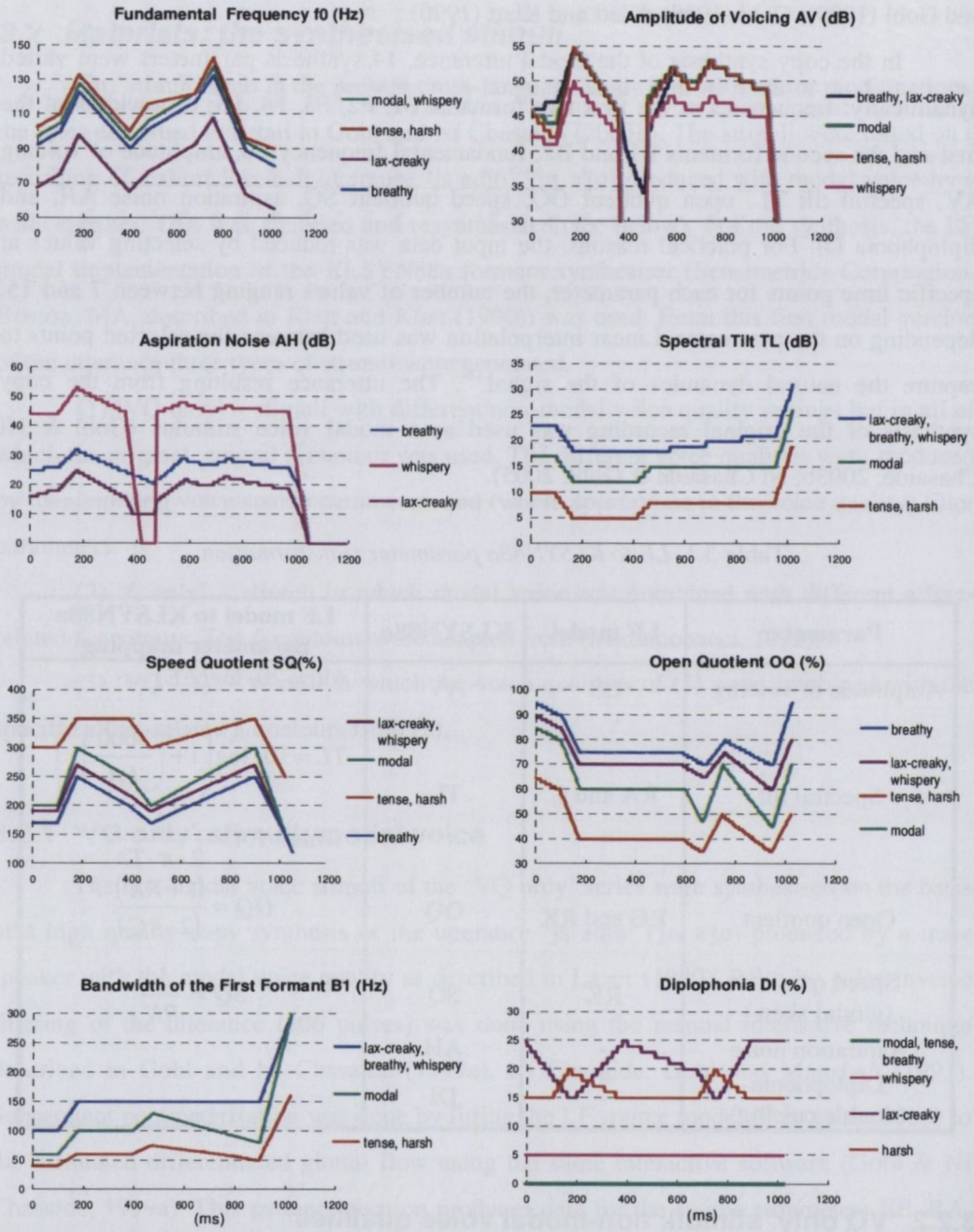


Figure 3.1. *Parameter variation in the synthesised voice quality stimuli (after Gobl and Ní Chasaide, 2003).*

quotient OQ, speed quotient SQ, aspiration noise AH, diplophonia DI) and the bandwidth of the first formant B1 (Figure 3.1). The dynamic variation of parameters was determined by earlier studies (Gobl, 1988; Gobl & Ní Chasaide, 1992; Ní Chasaide & Gobl, 1995) and was adjusted on the basis of the auditory analysis of the voice qualities (Gobl & Ní Chasaide, 2003b).

As evident from the parameter trajectories in Figure 3.1, non-modal voice qualities differed from the modal one in a number of settings. Tense voice has lower OQ, higher SQ, lower TL, narrower B1 and slightly higher f_0 values. Harsh voice used the settings for tense voice, with the DI parameter additionally manipulated to introduce aperiodicity of the signal. Breathy voice, relative to modal, has lower AV (adjusted on the basis of auditory judgment), higher OQ, lower SQ, higher TL, and wider B1. Lax-creaky voice used the source settings for the breathy voice stimulus with AH lowered by 20 dB relative to breathy voice, with the OQ values set to those of creaky voice, and f_0 lowered by 30 Hz compared to modal voice. As pointed out in Gobl and Ní Chasaide (2003b), Ní Chasaide and Gobl (2002), Ní Chasaide and Gobl (2005) this rather lax quality is different from the definition in Laver (1980) where creaky voice involves considerable glottal tension. Since in auditory impressionistic terms creaky voice may often sound quite lax, lax-creaky voice setting was considered plausible. Whispery voice was also based on the settings for the breathy voice with lowered AV, increased AH, slightly lowered OQ and slightly increased SQ (Gobl & Ní Chasaide, 2003b).

The synthesised voice qualities differed in terms of intrinsic f_0 values. For example, f_0 is 5 Hz higher for harsh and tense voice and 5 Hz lower for breathy voice compared with modal voice. This intrinsic f_0 differs most for the lax-creaky voice. The lax-creaky voice has intrinsically low f_0, and it was considered important to retain this feature in synthesis. As a consequence, its f_0 is 30 Hz lower than that of the modal (neutral) stimulus.

3.2.3 'f_0 only' stimuli

As mentioned earlier, the principal question relating to f_0 in this study was whether the addition of major dynamic f_0 contour differences to the non-modal voice qualities would enhance their affective ratings, particularly, for strong emotions. These combined stimuli are described in the next section. It was felt important to include a series where the f_0 contours were presented with modal voice ('f_0 only'), as a necessary control to allow differentiation of whether the combination of voice quality with f_0 perturbations or the latter on their own might account for any affective colouring found. These f_0 only stimuli are described below.

The contours used for the generation of 'f_0 only' stimuli were adapted from Mozziconacci (1995), see also Figure 3.2. The contours have two pitch accents and were considered appropriate as the shape of the f_0 contours is basically characteristic of declaratives in all four languages selected for this study. As evident from Figure 3.2, the f_0 contours differ, not in shape, but in terms of the level, range and dynamics: these are the kinds of features generally reported in studies on f_0 and emotion (see Chapter 2, Section 2.3.3 on the role of f_0).

In Mozziconacci (1995), quantitative data for seven affect-related f_0 contours were obtained by analysis of a short semantically neutral utterance containing two accented words pronounced by a male Dutch speaker who had been asked to portray indignation, anger, joy, fear, boredom, and sadness as well as a neutral state. In each of the emotional portrayals, f_0 was measured at six anchor points in the utterance corresponding to onset, two peaks, two values after the first peak and before the second peak, and offset, and averaged across three repetitions of the utterance. For all emotions, the shape of f_0 contour was found to be characterised by f_0 rise and fall on a single accented syllable, with the major difference being in the magnitude of the f_0 excursions and the rate of f_0 fall after the second peak (Figure 3.2). This pattern[19] was also the most frequent pattern in the dataset analysed in Mozziconacci (1998) and the author concludes that it is possible to express different emotions on this pattern. The contours differ in terms of the f_0 level, range and dynamics, such as the relative height of the 1st and the 2nd peaks and the magnitude and steepness of the fall in the 2nd peak (particularly steep in 'indignation' and 'anger').

The contours in Mozziconacci (1995) are based on production data in Dutch, for a range of affects. Their potency in signalling these same affects was tested in perception experiments on Dutch subjects, and each contour is thus associated with the affect it describes. Listening tests reported in the study showed better than chance emotion recognition. As in other studies, the author points out that pitch manipulation alone is not sufficient and other prosodic aspects such as duration and voice quality must be needed for the effective cueing of emotion (Mozziconacci, 1995).

No assumption is made here, however, that these f_0 contours should necessarily be associated with those particular affective states by the speakers of the other language groups tested in the present study. To some extent, even though the contours are referred to

[19] '1&A' (rise-fall on a single accented syllable) in terms of the Dutch intonation grammar ('t Hart et al., 1990) The same basic f_0 contours were found in expressive speech in Banziger and Scherer (2005), Grandjean et al. (2006).

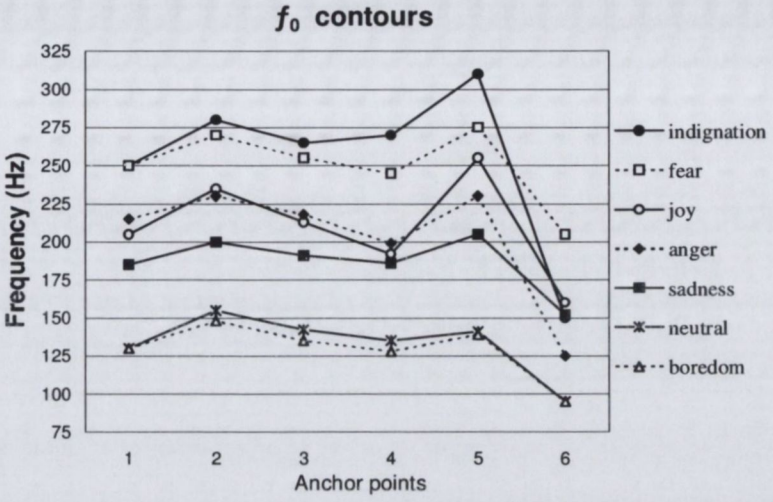


Figure 3.2. *Affect-related fundamental frequency contours (from Mozziconacci, 1995).*

in this thesis as f_0 'joy', f_0 'indignation' etc, we do not necessarily expect particular contours to have the affective valence found for Dutch. They can be regarded simply as set of f_0 contours that reflect likely affect-related variations, and which encapsulate the kinds of shifts in f_0 level, range and dynamics that are often mentioned in production based studies.

Thus, boredom and neutrality are characterised by low pitch level and small pitch excursions. Fear and sadness are represented by higher pitch levels, particularly high for fear which has a somewhat increased dynamic range. Anger, indignation and joy are all characterised by raised f_0 levels and an extended dynamic range, particularly in the second peak. Indignation is striking in having the very highest pitch level and a dynamically rapidly falling f_0 in the final accent. All in all, this seems to be a good sampling of the kinds of variations one is likely to get with this particular f_0 pattern.

As pointed out in Mozziconacci (1998), an emotion can be realised with different f_0 patterns and the same pattern can be used for different emotions. The shape of the contour proper, the A&1 pattern of the IPO intonation labelling system ('t Hart et al., 1990), is not necessarily affect-related, but this is the most likely default pattern of pitch movement and the one she would most readily recommend to use for affect-related manipulations. It is the contour dynamics, in particular in the realisation of the second peak, as well as the overall f_0 level and range that might contribute to affective loading. The contours should be viewed as such – f_0 contours of similar shape but with different magnitude of f_0 excursions (f_0 level and f_0 range) and different dynamics.

An approach alternative to adopting f_0 contours from prior work could involve systematic f_0 manipulation: e.g., the magnitude of f_0 excursions, relative height of the 1st and 2nd peak as well as the steepness of the declination line and the rate of change of f_0 in the second peak. This approach in itself would have many pitfalls, particularly concerning ecological validity: as the starting point would not be based on natural data one would be likely to produce some odd sounding output. In should be noted that in the present study no attempt is made to give full coverage to how f_0 cues affect, but rather I hope to get insights into its contribution relative to the voice quality cues. This study is not about the role of f_0 in affect signalling, rather it asks the question of whether extreme f_0 components are crucial, along with voice quality, in the signalling of the stronger emotional states, and if so, whether languages differ in this respect.

The use of the set of Mozziconacci (1998) contours was considered acceptable because (1) the set was obtained experimentally and tested for the presence of affective colouring in the course of listening tests; (2) the set presents a number of contours of the same overall shape but different in terms of the level, range and dynamics: these are the kinds of features generally reported in studies on f_0 and emotion (see Chapter 2, Section 2.3.3 on the role of f_0); (3) the shape of the f_0 contours is basically characteristic of declaratives in all four languages selected for this study. Thus, using these contours, it was considered possible to get insights into the role of f_0 when combined with the synthesised voice qualities.

This does not mean that the choice of f_0 contours is necessarily optimal. Note that the present f_0 contours are only one subset of possible contours, and that the differentiation (from modal) is sometimes small (and consequently, unlikely to have much effect. In particular, (1) there is very little difference between f_0 'neutral' and 'boredom', and so no difference in ratings between modal and modal + f_0 'boredom' could be expected; and (2) the 'sad' contour is higher than the neutral, whereas many studies mention f_0 lowering for sadness.

'f_0 only' stimuli were synthesised on the basis of the modal voice stimulus. Five f_0 contours - indignation, joy, fear, boredom, and sadness - were used in the synthesis of 'f_0 only' stimuli by a proportional scaling of the values in Mozziconacci (1995). As the f_0 contour of the original modal utterance was very similar to the neutral contour in Mozziconacci (1995), it was used as a neutral reference f_0 contour. The Mozziconacci f_0 values were scaled relative to this modal (neutral) reference at each of the anchor points so as to retain the relative differences with respect to the f_0 values to the modal (neutral) stimulus. Four additional anchor points were used for scaling: one before the 1st peak, one

between each peak and the lowest point between the two peaks, one after the second peak. The 'f_0 only' stimuli were synthesised by manipulating the F0 parameter in the KLSYN88a synthesiser. Figure 3.3 shows the relative difference between the f_0 values inherent for the 'VQ only' stimuli and the affect-related f_0 contours adapted from Mozziconacci (1995) and used in the generation of the 'f_0 only' as well as the 'VQ + f_0' stimuli (described below).

3.2.4 'VQ + f_0' stimuli

The stimuli of the third type were generated by combining non-modal voice qualities with the affect-related f_0 contours used in the synthesis of 'f_0 only' stimuli. The combinations of a particular voice quality with a particular affect-related f_0 contour were guided by the results of the earlier experiments as well as by comments in the literature (Gobl et al., 2002; Gobl & Ní Chasaide, 2003b; Laver, 1980; Murray & Arnott, 1993; Scherer, 1986). Thus, whispery voice was combined with f_0 'fear', breathy voice with f_0 'sadness', lax-creaky voice with f_0 'boredom', and tense voice with f_0 'joy' and f_0 'indignation'.

In combining particular voice qualities with f_0 contours the research aimed at testing whether the affective colouring of a particular voice would be enhanced by the dynamic f_0 contour, in particular for strong emotions. It has been shown experimentally that voice alone can impart affective colouring for mild affective states (Gobl and Ní Chasaide 2003), but the hypothesis here is that f_0 variation is needed for signalling stronger (full-blown) emotions.

Table 3.2 gives the summary of the stimuli used in the perception experiments of the cross-cultural study reported in Chapter 4. The stimuli are grouped according to the hypothetical 'target emotion' they are deemed most likely to signal, and there are stimuli of each type ('VQ only', 'f_0 only', 'VQ + f_0') within one such 'target emotion group'. Modal voice is used as a baseline reference for both 'VQ only' series and 'f_0 only' series (neutral f_0 contour). Note that the 'target emotion group' is essentially a way of grouping the stimuli for statistical testing into groups that have common elements. For example, the group 'sadness' includes the 'VQ only' stimulus with breathy voice, the f_0 contour associated in Mozziconacci (1995) work with sadness and the combined breathy + f_0 'sadness' stimulus. This grouping does not necessarily imply that there is a major expectation that the members of that group will yield the highest ratings for the particular emotion, e.g., sadness in this case.

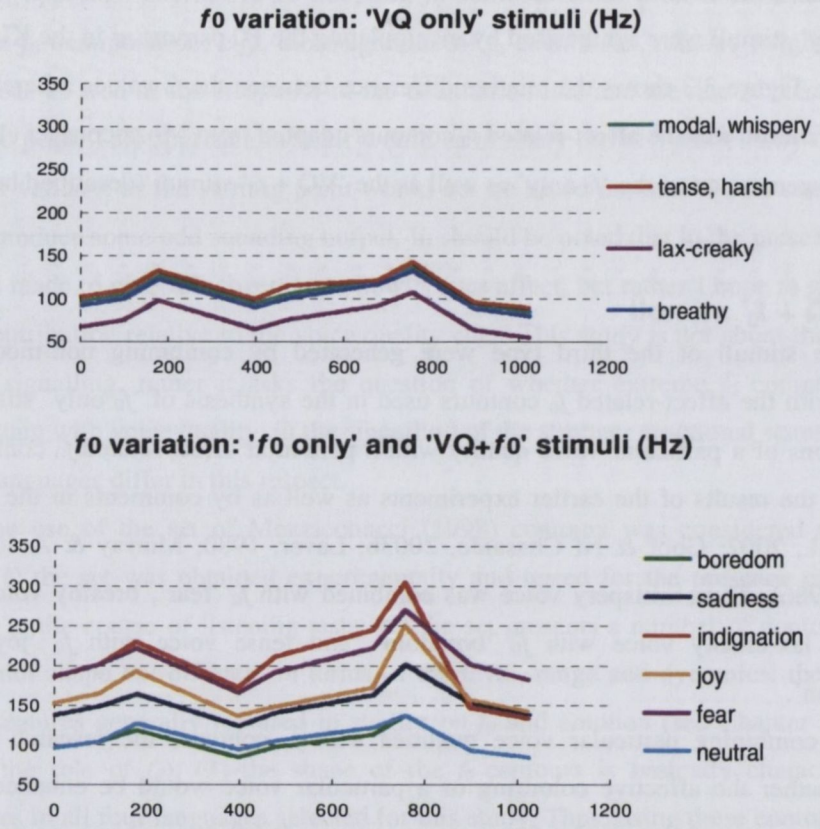


Figure 3.3. *Intrinsic to voice quality f_0 variation used in 'VQ only' stimuli and affect-related f_0 contours scaled after Mozziconacci (1995) and used in the synthesis of 'f_0 only' and 'VQ+ f_0' stimuli.*

Table 3.2. *Synthesised stimuli used in the cross-language study in this thesis.*

'Target emotion group'	**Stimulus type (type of manipulation)**		
	'VQ only'	**'f_0 only'**	**'VQ + f_0'**
fear	whispery	modal + f_0 'fear'	whispery + f_0 'fear'
sadness	breathy	modal + f_0 'sadness'	breathy + f_0 'sadness'
boredom	lax-creaky	modal + f_0 'boredom'	lax-creaky + f_0 'boredom'
joy	tense	modal + f_0 'joy'	tense + f_0 'joy'
indignation	tense	modal + f_0 'indignation'	tense + f_0 'indignation'
neutral	modal	neutral (= modal)	

Note that the range of the voice quality stimuli as well as the coverage of the f_0 contours and the possible f_0 and voice quality combinations is not exhaustive. The stimuli represent but a sample of possible f_0 contours, voice qualities and their combinations.

This thesis extends the prior work in that (1) in the experiments reported here all three sets of stimuli have been used and (2) the stimuli are rated by listeners from different language groups to test for the presence of universal/language-specific patterns in the use of voice and f_0 in affect signalling. As in the earlier studies, the listeners are presented with a range of stimuli, and are asked to rate them in terms of a fairly broad selection of affective attributes. These affective attributes include emotions (*scared, happy, angry, sad*) as well as milder affective states, attitudes and interpersonal stances (*relaxed, stressed, bored*; *formal, interested*).

3.3 Method: the perception tests

The 15 stimuli (see Table 3.2) were used in the perception experiments conducted according to the procedure described in Gobl and Ní Chasaide (2003b), as a series of six subtests. In each subtest, 10 randomisations of the 15 stimuli were presented to the participants and responses were obtained for a pair of opposite affective attributes (e.g., *sad-happy*). The pairs of affective attributes tested were *apologetic-indignant, bored-interested, intimate-formal, relaxed-stressed, sad-happy* and *scared-fearless.* A similar approach where stimuli were rated for affective content on a number of bipolar scales was used, for example, in Uldall (1964), Ladd, Silverman, Tolkmitt, Bergmann and Scherer (1985).

The stimuli were presented to the listeners in a quiet room through a high quality speaker with the volume set at the level considered comfortable listening level. Each subtest took approximately 11 minutes. The subjects had a short training session before the tests, and were given breaks between the subtests.

3.3.1 Scales

The participants were asked to judge the stimuli on six bipolar scales defined with contrastive adjectives at each end. This kind of semantic differential scale is commonly used in the study of attitude (Heise, 1970; Osgood, May, & Miron, 1975; Russell & Carroll, 1999; Streiner & Norman, 2008) and allows one to measure directionality of reaction (e.g., sad vs. happy[20]) as well as intensity (slight to extreme). The scale is usually interpreted as a 7 point scale where the neutral attitude (or in our case, 'no affective colouring') is assigned the value of zero (Heise, 1970, p.235). The same use of scale for measuring attitude in intonation contours is found in Uldall (1964).

The participants were instructed to judge each stimulus for the presence and strength of affect, and to mark their response on the answer sheet where the opposite affective labels were placed on each side with seven boxes in between. The choice of the centre box implied that no affective colouring was present in the utterance; checking the boxes to the left or right to the centre box indicated the presence and strength of a particular affect, the most extreme ratings being further from the centre box.

bored | | | | | | | interested

[20] Any pair of valenced antonyms can be used as opposite points of such scale (Russell & Carroll, 1999).

The scale was later interpreted as ranging from – 3 to + 3, where 0 value corresponded to no affect perceived, +/- 1 corresponded to mild affective colouring, +/-2 – to moderate and +/-3 - strong affective colouring. Numeric values were not presented to the raters during the experiment.

bored								interested
	-3	-2	-1	0	1	2	3	
	strong	moderate	mild	no affect	mild	moderate	strong	

3.3.2 Affective labels

The affective labels defining the opposite ends of each of the six scales have been chosen to cover a fairly broad range of emotions and milder affective states such as attitudes and interpersonal stances. The affective adjectives used in the scales are among those most frequently found in the lists of emotion-related words (e.g., Baron-Cohen, 2007; Douglas-Cowie et al., 2006; Juslin & Laukka, 2003). The choice of affective labels was in part guided by the synthesised voice qualities and by what is known about voice to affect mapping in the literature. The scales *relaxed-stressed, sad-happy, bored-interested, intimate-formal,* were used in prior work and were adopted here in part to assure comparability of results[21].

TRANSLATION OF LABELS

In the cross-language study reported in Chapter 4, the stimuli were presented to speakers of four different languages - Hiberno-English, Russian, Spanish and Japanese. For the speakers of languages other than English the affective labels on the answer sheets were presented in their respective languages in translation. The translation of the labels from English was undertaken by at least two native speakers of the respective languages (university students and staff) who had a good command of English and who were also familiar with the nature of the research and the purpose of the scales. As there are usually several translation possibilities, the translators were asked to discuss them and come to consensus regarding the best possible choice in order to keep the translation accurate and to maintain the polarity of the affective labels for the use in the scales. The affective labels and their translation are shown in Table 3.3.

[21] The scales used in Gobl and Ní Chasaide (2003b) *are relaxed/stressed, content/angry, friendly/hostile, sad/happy, bored/interested, intimate/formal, timid/confident* and *afraid/unafraid.* The scales used in Gobl et al. (2002) are *relaxed/stressed, content/angry, sad/happy, bored/interested, intimate/formal, afraid/unafraid,* and *indignant.*

Table 3.3. *Affective labels used in the perception experiments.*

English	Russian	Spanish	Japanese
apologetic	извинение	arrepentimiento	すまなさそう
indignant	возмущение	indignación	憤った
bored	скука	aburrimiento	退屈な
interested	интерес	interés	おもしろそう
scared	испуг	susto	おびえた
fearless	бесстрашие	audacia	大胆な
intimate	интимность	familiaridad	親しげな
formal	официальность	formalidad	かしこまった
relaxed	расслабленность	relajo	リラックスした
stressed	напряжение	estrés	ストレスがある
sad	грусть	tristeza	悲しい
happy	веселье	alegría	うれしい

Achieving conceptual/semantic equivalence of the scale anchor terms in translation is crucial for the interpretation of results. It is well documented in the literature, however, that emotion terms in one language do not map perfectly onto terms in the language spoken in different cultures (Mesquita et al., 1997; Ogarkova et al., 2009; Russell, 1991). In the extreme case a particular term in one language may even not exist in another language at all. For example, in Irish (Gaelic) there is no word that can be a direct translation of the English word 'bored'. In both Japanese and Spanish, the loan word from English 'stressed' is used for the affect, suggesting that it may be of recent origin in these languages/cultures, even though it is widely used today.

Although certain precautions were taken to render the translations as accurately as possible (team translation or the 'committee approach' that also involved back translation as recommended, e.g., in Brislin (1980), Streiner and Norman (2003, 2008), the non-equivalent of terms is a distinct possibility.

One case came to our attention where there was not a one to one equivalence. It concerned the words used to translate 'intimate' in Spanish. The translators chose the translation 'familiaridad' over what at face value might look like the most appropriate term 'intimidad'. This was discussed at some length. The translators did point out that the term 'familiaridad' did not include the sexual overtones of the English word 'intimate', they were nonetheless firmly of the opinion that 'familiaridad' was overall the closest in correspondence to the English term, and the appropriate polar opposite to 'formal'. The fact remains, however, that as there are differences in the semantic field of the affect, there might well be some differences in the voice used to express it.

It is understood that such differences in the semantic scopes of the affective labels has certain consequences for the data interpretation. In particular, if cross-cultural differences emerge in the data for such cases, the question inevitably will arise as to whether these differences are truly differences in the expression of affect, or whether they might (at least partly) reflect the fact that the affect being expressed is somewhat different in any case. Another issue that has been raised in this kind of research is whether translators themselves, who are usually better educated than the future participants might not go for labels that would be less frequently used by the subjects who take the tests (Streiner & Norman, 2008).

Taking the above considerations into account, it could be argued, however, that (1) the translation involved isolated words, and not complex multi-item constructions as is often the case; (2) the tests were administered to university level students, not a very different population from those who did the translation. Moreover, the subjects were given the opportunity to ask questions about the experimental procedure, this included questions about the meaning of the labels.

Very major cross language differences did emerge in our results (see description of results in Chapter 4), in particular, for the tests *intimate-formal* (for the Japanese and Spanish subjects) and *stressed-relaxed* (for the Japanese subjects). The author is fluent in the three European languages of this study, and in consultation with native speakers it was reasonably clear that, other than the issue with the term 'intimate' in Spanish, there was a rather good correspondence with the terms used. This was more difficult to ascertain for Japanese, and so a post-hoc test was conducted for this language to get further clarification on whether the interpretation of the scale by the participants might have been influenced by the translation of the labels. The test was conducted in order to identify the main points of possible discrepancy with the original translation that we hoped would help us with the interpretation of the results of the listening tests.

THE POST HOC TEST

Five Japanese speakers with a good command of English were asked to translate the labels used in the experiments reported here from English into Japanese. Five different speakers of Japanese were asked to translate the earlier versions of the translated Japanese labels back into English. The translators were not aware of the nature of the scales or of the aim of the translation, they did the translation individually and no panel discussion was held to arrive at the best possible variants. Although the subjects were asked to provide one translation label, in certain cases they offered several alternative translations. The results are given in Tables 3.4 and 3.5. The interpretation of results was done in consultation with

a native speaker of Japanese who was familiar with the nature of the research reported in this thesis and who has a good command of English and who has lived in English speaking countries for a number of years.

THE ENGLISH→ JAPANESE TRANSLATION

In the English → Japanese translation, the following differences with the original translation were identified (they are shaded in Table 3.4):

1) 'scared' was translated into Japanese as 'scary';

2) 'fearless' was translated as both 'bold' and 'fearless';

3) 'look/sound intimate' in the original translation vs. 'being intimate' in the post-hoc test translation;

4) a different term was chosen to represent formality. The translators in the post-hoc test chose the term 'keishikibatta' (形式張った), which is a literal translation of the word 'formal' ('form' + verb 張る 'to pose/enforce the form'). The term can be applied not only to the attitude of people, but also to the style of things as well. However, it is less likely to represent affect, whereas the original term 'kashikomatta' (かしこまった) has more to do with psychological distance and is more related to the affect/attitude of being formal.

5) 'happy' was translated as both 'happy' and 'glad'.

With the exception of *formal*, these differences between the original translation and the one provided in the course of the post hoc test are relatively minor. In discussions following the test it was pointed out by Japanese speakers that in the case of 1 and 3 one is dealing not so much with semantic differences as with a grammatical issue. In Japanese, the same adjective may mean 'boring' and 'bored', and a broader context, e.g., an auxiliary verb is required to convey the exact lexical meaning. Note that in this test the translators dealt with isolated words and were not aware that these form a part of a bipolar scale, but subjects pointed out afterwards that the bipolar context of the test would clarify and elicit the intended interpretations.

The translators were uniform in the choice of the Japanese words for *stressed-relaxed*. Note that the perception test showed a surprising lack of affective response from the Japanese subjects unlike the response from the three European language groups of the study. This difference, we conclude, is unlikely to be attributable to the interpretation difficulties the subjects may have had with the labels defining the scale.

Table 3.4. *Translation from English into Japanese. The original translation is shown in the boxes with a thicker black border. Shaded table cells indicate differences in translation.*

S	Apologetic	Indignant	Bored	Interested	Intimate	Formal	Relaxed	Stressed	Sad	Happy	Scared	Fearless
	すまなさそう（な）	憤った	退屈な	おもしろそう（な）	親しげな	かしこまった	リラックスした	ストレスがある	悲しい	うれしい	おびえた	大胆な
1	すまなそうな	憤慨した	飽きて	面白そうに	親密な	形式張った	リラックスした	ストレスがある	悲しい	嬉しい	おびえて	怖がらずに
2	すまなそうな	憤慨した	退屈な	興味のある	親密な	正式な	リラックスした	ストレスの溜まった	悲しい	嬉しい（うれしい）	怖い	大胆な
3	謝罪的な	憤慨した	退屈	興味ある	親密な	正式な	リラックスした	ストレスがある	悲しい	幸せな	怖い	恐れない
4	お詫びの	憤慨した	退屈した	興味のある	親しい	フォーマル	リラックスした	ストレスを感じる	悲しい	幸せ	怖がった	怖いもの知らず
5	すまなそうな	憤慨する	たいくつな、つまらない	興味がある、おもしろそうな	親密な	形式張った	のんびりした、落ち着いた	ストレスがある	かなしい、かわいそうな	うれしい	こわい、おそろしい	大胆な

Table 3.5. *Translation from Japanese into English. The original translation is shown in the boxes with a thicker black border. Shaded table cells indicate differences in translation.*

S	Apologetic	Indignant	Bored	Interested	Intimate	Formal	Relaxed	Stressed	Sad	Happy	Scared	Fearless
	すまなさそう（な）	憤った	退屈な	おもしろそう（な）	親しげな	かしこまった	リラックスした	ストレスがある	悲しい	うれしい	おびえた	大胆な
1	look sorry	furious	boring	look/ sound interesting	friendly	formal	relaxed	stressful	sad	glad	frightened	bold
2	sorry	mad	boring	interesting	friendly	formal	relaxed	stressed	sad	happy	scared	bold
3	look sorry	enraged	boring	look interesting	friendly	formal, textbook-like	relaxed	stressed	sad	happy	frightened, scared	bold, audacious
4	sorry, remorseful	angry, indignant	boring	interesting-sounding	friendly	1) formal, stiff; 2) respectful, humble	relaxed	stressed, stressed-out	sad, unhappy	happy, joyful	scared, frightened, terrified	bold
5	sorry, apologetic	outraged, furious	boring	interesting	affectionate, friendly, intimate, familiar	formal, reserved	relaxed, at ease	stressed (out)	sad, sorrowful	happy, joyful	fearful, scared	bold, fearless

The term 'formal' was translated originally by a term suggesting 'psychological distance' while the translators in the post hoc test selected a literal kind of translation. This will be further discussed below.

THE JAPANESE → ENGLISH TRANSLATION

In the Japanese → English translation, the main differences were as follows (they are shaded in Table 3.5):

1) 'bored'→ 'boring',
2) 'interested' → 'interesting',
3) 'intimate' → 'friendly'.

The former two are only minor points of discrepancy between the original translation related to the fact that in Japanese the same adjective is used to mean 'interested/interesting' or 'boring/bored' (or, as in the above, 'scary/scared'). The translation of 'intimate' in the post-hoc test was also consistent with the original translation. However, it requires further discussion (see below).

DISCUSSION AND CONCLUSIONS

Following the implementation of the test, I discussed the labels with some native speakers, and I will summarise the results and conclusions for the four affects which as mentioned above yielded striking cross language differences *stressed, relaxed, intimate* and *formal.*

Stressed: There did not seem to be any difficulty with the label used here, or with the way in which the label was employed. The subjects of the post-hoc test used the same term. Native speakers also assured me that the term is used in a very similar way to the English term.

Relaxed: Similarly, there did not seem to be any labelling/translation issue with this term.

Intimate: In the translation into Japanese, the same term was chosen as in the perception test of this study. In translating from Japanese to English, there was a predominance of *friendly,* although the terms *intimate, affectionate, familiar* also arose. In an experimental study of the conceptualisation and expression of intimacy by Japanese and Americans reported in Seki, Matsumoto and Imahori (2002), the most common translation of the word *intimacy* by the majority of the Japanese participants corresponds to our original translation of the label *intimate*. In that study, the word 'intimacy' was translated as shinmitsu-sa = 親密 (shinmitsu) -さ (sa = 'a degree of') or shitashi-sa = 親し(shitashi)

-さ (sa = 'a degree of'). The translation adopted here is 親し-げ-な (shitashi-ge-na) = 親し ('intimate/familiar') + げ (ge = 'look/sound') + な (na-adjective morpheme). Four of the five subjects in the post-hoc translation test translated the label 'intimate' as 親密 (shinmitsu) + な (na-adjective morpheme) and one translated it as 親し (shitashi) + い (i-adjective: morpheme) (Mika Ito, personal communication). It can be concluded that the original translation corresponds with the general view of the way this label can be translated.

Following the test, in discussion about the range of meanings of the Japanese and English terms, it emerged that there is in fact some difference in the semantic scope of the labels in the two languages. While 'intimate' in English includes 'close, familiar' as well as including possible sexual overtones, in Japanese the meaning is 'close, friendly, familiar', but without the sexual overtones. So, although we can be fairly confident that the term used in the present perception study is the best equivalence for Japanese, we also note that there is also some difference in the semantic field of the term, and that this might be a contributory factor in the very extensive differences that emerged in the perceptual results. Interestingly, the semantic differences here are rather like the differences pertaining to the Spanish them used for intimacy, in that the Spanish term also does not include the possible connotation of sexual intimacy. It is also interesting that in this test these two languages were broadly similar in their responses.

Formal: The term chosen used in the present perception study turned out not to be the one favoured by the subjects in the post-hoc test. In discussing this following the test, it appeared to be the case that the term used in the present study (かしこまった 'psychological distance') is a term more associated with human affect and attitude and therefore more suitable for the purposes of our study than the one chosen in the post-hoc test. The literal translation of the word 'formal' in the post-hoc test can be associated with a style which is not necessarily related to human attitude. Note that while the original translation was done bearing in mind the purpose of the cross-language experiment and the bipolarity of the scale, the participants in the post-hoc test dealt with isolated words. It was further pointed out by Japanese speakers that the original translation was not at all a complicated literary or high-brow term. We tend to conclude that (1) the term used (かしこまった 'formal, psychologically distant') indeed corresponded to the intended meaning in English; (2) it is unlikely that the original translation could have caused interpretation difficulties for the Japanese participants in the cross-language study; it is therefore unlikely that the original translation could have contributed to the results in any significant way. Undoubtedly, there

are likely to be other factors that might explain the very different results. In particular the *keigo* system of the formal expression of politeness in Japanese is likely to be an important factor (see Chapter 4).

3.3.3 Participants

The participants were speakers of four different languages, Hiberno-English, Russian, Spanish and Japanese. As it was desirable to exclude the influence of semantic content of the Swedish utterance used as the basis for the synthesised stimuli on the perception of affective, we made sure prior to the experiments that none of the participants spoke Swedish. It is virtually impossible in the modern world to completely exclude the influence of foreign languages and with this the influence of language-specific 'foreign' voice qualities, but such an influence was kept to a minimum in the present experiments as none of the participants had lived in a foreign country for any length of time or used foreign languages in an active way in their professional life (e.g., as an interpreter or a language teacher).

There were minor differences in the number of participants in the groups. The details about the participants are outlined below.

Hiberno-English subjects: N = 20, 7 male and 13 female, aged 18-35; students of Trinity College Dublin, Ireland.

Russian subjects: N = 21, 11 male and 10 female, aged 18-35; students of St. Petersburg State University, Russia.

Spanish subjects: N = 20, 10 male and 10 female, aged 18-22; students of Universitat Pompeu Fabra, Barcelona, Spain.

Japanese subjects: N = 21, 11 male, 10 female, aged 18-22; students of University of Tokyo, and Ochanomizu Women's College, Tokyo, Japan.

The choice of the listeners for the cross-cultural study was in part guided by my own intuitions about the neutral voice quality used in these languages. My impression is that Hiberno-English and Russian neutral voice are relatively similar and that Spanish and Japanese are rather different in that respect.

Although the reaction of the listeners to the stimuli was not entirely predictable at the outset of the experiments, certain expectations could be gleaned from the literature on the use of voice quality and f_0 in the four languages.

The shape of the intonation contours (Figure 3.3) represents a more or less universal f_0 pattern that intonation descriptions of all four languages include (Abe, 1998; Alcoba & Murillo, 1998; Svetozarova, 1998; Wells, 1982).

As for the use of voice quality, Japanese is probably different in the most striking manner compared to the European languages. Breathy voice in particular has been studied extensively in Japanese, where it is related mainly to the expression of politeness in both male and (although to a lesser degree) female speakers, e.g., Ito (2005). Campbell (2004a) reported voice quality variation in a Japanese female speaker (primarily, the degree of breathiness and f_0 variation) depending on the addressee measured as NAQ (Alku et al., 2002) using a substantial naturalistic speech corpus recorded over a period of more than two years. He found, for example, that NAQ is highest (suggesting increased breathiness) when the speaker was addressing non-family members (expressing politeness) or children, whereas self-directed speech showed the lowest values for NAQ. Similar findings are also reported in Campbell and Mokhtari (2003). The role of whispery/breathy voice in expression of affect was studied in Ishi, Ishiguro and Hagita (2008) on the material of spontaneous dialogue speech of several speakers. Their results showed that breathy/whispery voice (the authors made no distinctions between the two voice qualities) could be used in the expression of emphasis, excitement, polite interest, gentleness or tenderness. A more whispered voice quality appears in confidential talking, expressions of embarrassment, or when the speaker is talking to oneself. According to a perceptual study of voice quality stereotypes of heroes and villains in Japanese animation (2004), breathy voice is associated with positive characteristics (voices of heroes had little laryngeal constriction), whereas tense and harsh voices and raised larynx tend to be associated with negative characteristics (stereotypical villains). Maekawa (1998) reported a study of prosodic characteristics of Japanese acted emotional speech. The emotions ('paralinguistic information types') included admiration, disappointment, suspicion, indifference, focused and neutral. Suspicion, admiration and disappointment were marked by the use of creaky voice ('laryngealisation'). According to Sadanobu (2004), who studied the use of *rikimi* (pressed/creaky voice) in Japanese in spontaneous speech in everyday interaction, pressed voice is used in Japanese in exclamation, expressions of surprise, irritation, disgust and anger. Listening tests, in which repetitions of the same sentence uttered with modal and pressed voice were compared, showed that pressed voice was associated with stronger expression of emotion than modal. Sadanobu (2004) identifies the use of pressed voice with suffering, physical or moral (e.g., lifting heavy objects or trying to solve a difficult problem; utterances of request are often uttered with pressed voice and a grimace), as well as with the expression of admiration and emphasis. He further comments on very subtle context-dependent use of *rikimi* in Japanese related to everyday communicative strategies. For example, the speaker needs to have the right to utter with pressed voice (it can only be

used when talking about one's own experience rather than abstract knowledge). Utterances of request and rejection sound more polite and sincere with pressed voice.

Raising the pitch of the voice has been reported as a means used by a male actor to express contempt in Japanese (e.g., Kramer, 1964). According to Abe (1998), Japanese women, and to a lesser degree males working in service industry use high pitched voice as a politeness formula or, in case of male service industry workers, as a way to show willingness to satisfy customers' wishes. High pitch in female speech has been reported to signal such socially attractive characteristics as cuteness, kindness and politeness whereas low pitch in females was judged as characterising the speaker as stubborn, selfish and strong (Ide & Yoshida, 1999). In Maekawa (1998), pitch range in Japanese was reported to be strongly compressed in disappointment, and enlarged in admiration, focused and suspicion. According to Ofuka, McKeown, Waterman and Roach (2000), pitch in Japanese changes mainly on the final vowel in the sentence, which is particularly important for expressing the speaker's state, since particles signalling attitudinal meanings are generally located at the end of a sentence.

The use of voice quality in expressive speech in Spanish has been studied in particular with relation to emotional speech synthesis (Montero, Gutiérrez Arriola, Colás, Enríquez, & Pardo, 1999; Montero Martínez, Gutiérrez Arriola, de Córdoba Herralde, Enríquez Carrasco, & Pardo Muñoz, 2002). Montero et al. (1999; 2002) presented analysis of the Spanish Emotional Speech Database that included simulation of happiness, sadness, cold anger, surprise, as well as neutral, by a male professional actor. The corpus was evaluated in a series of listening tests using copy diphone synthesis of mixed type (e.g., diphones could be taken from the neutral sentence and its prosody modified as in the happy one). Their results showed, for example, that surprise had the highest mean f_0; cold anger was characterised by the presence of 'voice source noise' as well as a considerable amount of vibrato/tremolo; sadness was characterised by a flat low f_0 and slow rhythm. A more detailed analysis of voice quality in affective speech in Spanish is reported in Monzo, Alías, Iriondo, Gonzalvo and Planet (2007) who focused on the capability of voice quality parameters to differentiate among a number of expressive speech styles (neutral, happy, sad, sensual and aggressive) using an expressive speech corpus in Spanish (female speaker). They found that each of these styles was associated with a specific voice quality, e.g., neutral speech style was characterised by modal phonation; happy - by 'mid-harsh' phonation; sad – by 'mid-whispery' phonation; sensual – by whispery phonation; aggressive – by harsh phonation. The auditory impression that Spanish female speakers habitually use somewhat tenser voice quality than, for example, American English

speakers is supported by findings of the acoustic analysis reported in Trittin and de Santos y Lleó (1995).

Practically no studies of the use of voice quality and f_0 in Russian in affect communication have been found to date. Intonation of positive emotions (interest, surprise, joy) has been compared for Russian and English spontaneous speech in Vysotskaya and Kovyazina (2002) in terms of pitch level, pitch range and intensity level and showed both similarities and cross-language differences. Listening tests in which Russian listeners were asked to judge the emotional load of utterances pronounced by English speakers showed that joy was best recognised, followed by interest; recognition rates for surprise were quite poor. As for the voice quality, the publications on this matter are indeed scarce, but see, for example, Andreeva and Krivnova (2005) on the use of laryngealisation in Russian dialogue speech.

Overall, the experimental evidence points at interesting distinctions in the use of voice quality in the languages/cultures selected for the analysis. Although no prior assumptions regarding cross-cultural similarities/differences in stimulus to affect association could be made, certain culture-specific patterns were expected to emerge in the course of perception tests.

3.4 Scope and limitations of the study

Before describing the statistical procedures used in the experiments it is important to outline the scope and limitations of this study as well as to point out certain methodological issues that might pose constraints on data interpretation and generalisation.

First, the selection of voice qualities used is not exhaustive. Although an attempt had been made to include a broad sampling of voice qualities, there are inevitably omissions, e.g., falsetto, which has sometimes been found to be potent in signalling fear. It is possible that the omissions have potentially more serious consequences for one language group than for another. There could be a frequently used voice quality, say, in Japanese that does not particularly feature in English or vice versa.

Voice qualities are not represented as their extreme exemplars: it is possible to have many degrees of a particular quality. It was shown by Ryan, Ní Chasaide and Gobl (2003), where voice source parameters were varied to produce tense to lax continuum, that the strength of affective responses covaried in a more or less continuous way.

These points need to be borne in mind when interpreting the results. For example, if the voice quality stimuli of the present study do not elicit a particular affect, we cannot conclude that this affect is not signalled by voice quality. In other words, we cannot ignore the possibility that a quality not included, or a more extreme version of a quality that is included might have elicited an affective response.

Similar points need to be made in connections with the 'f_0 only' stimuli and the combinations in 'VQ + f_0' series.

The 'f_0 only' stimuli were also a very limited (but hopefully representative) sampling of the possibilities that may exist, including that of varying the f_0 contour. Again, the results need to be interpreted in this light. If a particular f_0 contour is found to elicit a high rating for a particular affect, this can probably be trusted, if however the f_0 contours provided in this study fail to elicit a particular affect we cannot conclude that f_0 does not play a role, but rather that these particular f_0 contours are ineffective.

In the case of 'VQ + f_0' contours the same kinds of limitations pertain. There are two further issues. First of all, the particular combinations were made on the basis of likelihood. In other words, they were guided by the literature and by the results of the earlier experiments (Banse & Scherer, 1996; Gobl et al., 2002; Gobl & Ní Chasaide, 2003b).There are inevitably going to be gaps in the coverage of these combined stimuli. Furthermore, the combined stimuli are also likely to be somewhat English-centric, given that the earlier experiments and most of the literature that guided these combinations are

based on English. There is at least a theoretical possibility that the most appropriate ways of combining voice quality and f_0 differ for different languages. Therefore, in interpreting results the same note of caution is needed. If a particular combination is found to potently evoke an affect, we can probably trust that result. If a particular affect fails to be evoked by any of the combined stimuli, we cannot conclude that there is not some other combination that might successfully do so.

The voice quality stimuli were generated based on previous analyses of different voice qualities and not on a prior investigation of affective speech. The interpretation of results and the inferences about the use of voice and f_0 in speech communication can only be made as far as the quality and naturalness of the synthesised stimuli and the quality of the f_0 contours allow. The naturalness of the stimuli was assessed informally by a phonetician trained in the Laver's framework as well as by expert listeners (Gobl and Ní Chasaide) during their work on the synthesis. Note that in the course of the listening tests (discussed in Chapter 4) the listeners were unaware that the stimuli were synthesised, and none of them raised the issue of the naturalness of the samples they listened to. In future work, formal assessment of the naturalness of the synthesised stimuli is desirable.

Further possible limitations that need to be borne in mind in the interpretation of results pertain to the selection of affective labels that define the rating scales. The affective labels have been chosen to cover a fairly broad range of emotions and milder affective states such as attitudes and interpersonal stances. The labels do not represent, however, equally easily interpretable states. For example, affective states such as *apologetic, intimate, formal* involve considerable cognitive component and require a different kind of perceptual effort compared *to sad, happy, bored*, and *scared.* There is also a possibility that the scales be interpreted very differently cross-culturally due to the lack of semantic equivalence of the scale anchor terms in translation.

The scales were constructed so that to maintain the polarity of affective labels. However, some affective states are not as clearly bipolar as others (e.g., *sad-happy, bored-interested, stressed-relaxed*, are intuitively bipolar, whereas *apologetic-indignant, scared-fearless,* and *intimate-formal* are less so). Furthermore, the 7 point bipolar scale used here has also been widely used in the study of attitude; however, it might appear rather coarse and allow for only fairly broad differentiation of affective colouring.

3.5 Research questions

The study focuses on the following research questions:

- How do 'VQ only' stimuli map to affect? How do these stimuli yield affective ratings? Does this mapping support traditional description found in the literature?
- Does the addition of a dynamic f_0 contour enhance affective rating of a particular voice quality stimulus? The hypothesis here is directional: the research is predicting that f_0-enhanced stimulus from the 'VQ + f_0' series will be more effective in signalling affect, particularly strong emotions, than the corresponding stimulus from the 'VQ only' series.
- To what extent can the f_0 contour (alone, i.e. with modal voice) induce the same affective rating?
- What stimuli (other than 'VQ only' stimuli) are consistently associated with a particular affect?
- Do the individual languages studied here associate voice quality (and f_0 contours) with affect in a similar way, or do distinct language-specific trends emerge?

3.6 Statistical analysis

3.6.1 Design

To allow for a three-way comparison among the ratings of the three types of stimuli, that is, the 'VQ only' stimuli relative to the 'VQ + f_0' and the 'f_0 only' stimuli, within each of the 'target emotion groups', taking the language factor and the grouping into account, a separate repeated measures ANOVA was conducted for each subtest using the SPSS software package. A complex mixed design was applied. The two within-subjects factors were the 'Target emotion group' (5 levels: fear, sadness, boredom, joy, indignation) and 'Stimulus type' (3 levels: 'f_0 only', 'VQ only', 'VQ + f_0'). The between-subjects factor was 'Language' (4 levels: Hiberno-English, Russian, Spanish, Japanese). The dependent measure was affective rating (averaged across 10 randomisations for each speaker).

As tense voice was used twice, in the generation of the two 'VQ + f_0' stimuli from two 'target emotion groups', joy (tense + f_0 'joy') and indignation (tense + f_0 'indignation'), it was not possible to apply the 5 x 3 x 4 factorial design to the data. In each subtest, the analysis of variance was therefore done twice as a 4 x 3 x 4 design (4 'target emotion groups' x 3 'stimulus types' x 4 'languages'), first for the stimuli in the 'target emotion

groups' fear, sadness, boredom and joy and second, for the stimuli in 'target emotion groups' fear, sadness, boredom and indignation.

To assess the ratings of 'VQ only' stimuli relative to modal and the ratings of 'f_0 only' and 'VQ + f_0' stimuli relative to modal (neutral f_0), a separate repeated measures ANOVA test of simple contrasts was conducted for each subtest, in which all the stimuli were compared to modal voice (neutral f_0) as it served as a baseline for both 'VQ only' and 'f_0 only' stimuli. The within-subject factor was 'Stimulus', the between-subject factor was 'Language'. The dependent variable was the affective ratings for each subject averaged across 10 randomisations. To correct for multiple comparisons and to minimise the risk of random effects, a Bonferroni correction was applied in the post-hoc tests.

The discussion of results will primarily focus on ratings above +/-1. Attempts have been made at data reduction by examining the mean ratings yielded by each stimulus for each group of listeners. To establish if the data would give some indications for such reduction, rank ordering of mean affective ratings was considered, but in the majority of cases there was no dramatic shifts among the ratings of different stimuli, the increase/decrease was incremental. Where such a shift was observed, it usually occurred at around +/- 0.9-1 on the rating scale. An admittedly arbitrary threshold of +/-1 was therefore adopted. This is not to mean that the ratings below this threshold are necessarily of no importance, and indeed statistically significant difference can be found between the ratings that are quite low. Rather, by examining and discussing the ratings above 1 we are more likely to focus on more robust and consistent voice to affect associations.

3.6.2 Interrater agreement

As outlined in Kreiman, Gerratt, Kempster, Erman and Berke (1993), there are multiple sources of variability in voice stimuli ratings, which include listener factors (sensitivity, bias, error, fatigue, etc.), scale factors (scale resolution and the way the scale is defined), stimulus factors (the quality of voice samples or synthesised stimuli), as well as the interaction of these factors. In an experiment such as reported here, absolute agreement would not be expected, but consistency in assignment of similar ratings to the same stimulus will suggest stronger voice to affect association.

To establish whether the listeners rated voice qualities in a coherent fashion, listeners' agreement/consistency in ratings was measured in a number of ways.

Single measures and average measures Intraclass Correlation Coefficients (ICC) (Landis & Koch, 1977; Shrout & Fleiss, 1979; Yaffe, 1998) were calculated for each

subtest for each language group of listeners and overall[22]. Since the stimuli used here represent only a sample of possible voice qualities and f_0 combinations, and since the listener judges were randomly selected from a larger population, the two-way random model was used (McGraw & Wong, 1996; Yaffe, 1998). As it is of interest to establish whether we can assume that the judgment of one rater is similar to that of the others, the single measures ICC (r) rather than the average measures ICC (R) will be mostly considered here as an indicator of raters' consistency. Following Landis and Koch (1977), ICC of 0.40-0.59 will be interpreted here as moderate interrater agreement, ICC of 0.60-0.79 will be interpreted as substantial interrater agreement, and ICC of 0.80 and above - as outstanding interrater agreement[23].

To identify the subjects whose ratings differ substantially from those of other raters, for each subtest and each language group the ratings of individual raters was correlated with the group mean. The results are given in Appendix 5.

Percent agreement was also used to establish the percentage of the raters that assigned relatively high affective ratings to individual voice stimuli as opposed to the ratings suggestive of no affective colouring. To calculate percent agreement, the data was examined and the mean ratings ranging from 0 to 1 were considered as not indicative of clear affect signalling (= 'no affect'). Mean ratings of over 1 were considered as indicative of confident affect attribution. The corresponding percentage of cases in which a certain voice quality was perceived as signalling affect or as having no affective colouring was calculated. The data is presented in Appendix 4.

The experimental design involved presenting each of the stimuli to the listeners 10 times in randomised order, and the listeners had to make very quick decisions regarding the association of stimuli and affects. The data for each listener across 10 randomisations was averaged, and all statistical analyses were done using this averaged measure. While random variation in the responses of an individual rater is highly likely, averaging across this relatively high number of randomisations was intended as a means to arrive at a robust trend for a particular listener. The random variation within the responses of a particular rater was not considered of particular interest for the purposes of this study as such, but it could contain information about the confidence of each individual rater in rating a particular stimulus. This can be inferred by inspecting the standard deviation of the mean across 10 randomisations for each listener, for each stimulus, in each subtest. The data is

[22] The measure is preferred to Pearson's correlation as a measure of the overall coherence of an entire group of listeners as it is less sensitive to the error inherent to the individual measures (Kreiman et al., 1993; Streiner & Norman, 2008).

[23] The terminology is that used in Landis and Koch (1977).

presented as scatter plots in Appendix 6. Given the range of the rating scale, the standard deviation of 0-1 would suggest a reasonably high degree of confidence in ratings while the standard deviation above 1 would point at an increased randomness in stimulus to affect association.

The results of the cross-language experiment are presented in Chapter 4.

4 Chapter four
Voice-to-affect mapping: cross-language study

4.1 Introduction

Chapter 4 discusses the results of the cross-language/culture study of voice-to-affect mapping, with speakers of four different languages – Hiberno-English, Russian, Spanish, and Japanese as participants. The main research question is to establish how particular voice qualities are associated with affective states in these languages and whether these associations can be enhanced by combining particular voice qualities with f_0 contours of varying dynamics. Data analysed here also provide insights into what might be universal in the way the voice and f_0 cues are used to communicate affect as well as to draw attention to major cross-language/cross-cultural differences.

The stimuli and research methodology are outlined in Chapter 3. As mentioned earlier in Chapter 3, the research focuses on the following questions:

- How do 'VQ only' stimuli map to affect? How do these stimuli yield affective ratings? Does this mapping support the traditional description found in the literature?
- Does the addition of a dynamic f_0 contour enhance the affective rating of a particular voice quality stimulus? The hypothesis here is directional: the research is predicting that f_0-enhanced stimulus from the 'VQ + f_0' series will be more effective in signalling affect, particularly strong emotions, than the corresponding stimulus from the 'VQ only' series.
- To what extent can the f_0 contour (alone, i.e. with modal voice) induce the same affective rating as the corresponding VQ+f_0 stimulus?
- What stimuli (other than 'VQ only' stimuli) are consistently associated with a particular affect?
- Do the individual languages studied here associate voice quality (and f_0 contours) with affect in a similar way, or do distinct language-specific trends emerge?

The discussion of results will therefore be done following these research questions. The results are presented in order for each subtest (pair of affects). In every case, there is a short initial summary of the statistical analyses, and following this the cross-language results for each affect are discussed in some detail. For each affect, I also summarise briefly any specific predictions in the literature for the languages in question.

The results are plotted in Figures 4.1 - 4.6. These show for all the stimuli (arranged in terms of the 'target emotion groups') the extent of the affective association, interpreted as the magnitude of rating – distance from 0. As mentioned in Chapter 3, the focus is primarily on ratings above +/-1, i.e. where the strength of the ratings is considerably in evidence. In Figures 4.1 - 4.6, the area below this arbitrary threshold (the area of no practical interest/low affective significance) is shown in grey. The cases deemed of particular interest for the discussion are marked by red rectangles.

In every case, it is of particular interest to ascertain whether the ratings are significantly different among the language groups. Statistically significant differences in affective ratings for a particular stimulus across languages are shown in Figures 4.1 - 4.6 with black asterisks and the letters E, R, S, J indicate for what language groups the difference in ratings was found statistically significant.

The discussion is done according to the following structure. Firstly, I consider which, if any, of the 'VQ only' stimuli used in this study yield highest ratings for the particular affect. It is of interest not only to ascertain which stimulus/stimuli impart(s) an affective colouring, but also to what extent the affective rating has shifted relative to the modal stimulus. Furthermore, it is relevant to see whether a particular affect is strongly signalled by a particular voice quality stimulus, or by more than one. It is also of interest to see whether traditional claims in the literature are supported.

Following this, I examined how the 'f_0 only' stimuli were rated, comparing them both to the neutral f_0 stimulus (modal) stimulus. It is of interest to see to what extent the present ratings match our expectations (if we had such).It is also of interest to compare the more highly rated 'f_0 only' stimuli with the 'VQ only' stimuli.

The combined stimuli 'VQ + f_0' are then examined, to see which combined stimulus yields the highest rating for the particular affect and also to consider how they 'perform' relative to the other two series of stimuli. For example, is the affective colouring of the most effective 'VQ only' stimulus enhanced by the addition of a non-neutral f_0 contour, as was initially predicted, for the strong emotions at least? It is also of interest to make comparisons with the corresponding 'f_0 only' stimuli.

The issue of universality or cross-language differentiation is then discussed, bringing together any trends that have appeared in response patterns for the different stimulus types, rated for the particular affect.

Finally, a short summary of findings concludes the discussion of results for each affect.

I recap briefly here on the statistical treatment which has been described fully in Chapter 3. In the first test, a complex 3 x 4 x 4 mixed design (3 stimulus types) x (4 target emotion group) x (4 language) was used to assess the effect of the type of stimulus ('VQ only', 'f_0 only', 'VQ + f_0') and the 'target emotion group' on the strength of affective ratings across the language groups tested. The reader is reminded that a 'target emotion group' includes a particular voice quality stimulus, a particular f_0 contour which could be associated with a particular emotion (e.g., joy), as well as a combination of the two (see Chapter 3). These two factors ('stimulus type' and 'target emotion group') were the within-subjects variables in the design. The between-subjects variable was the language of the listeners. The dependent variable was the affective rating. For each of the 6 subtests (*apologetic-indignant, bored-interested, intimate-formal, relaxed-stressed, sad-happy, scared-fearless*), the repeated measures ANOVA was conducted in two parts. This had to do with the fact that, due to the structure of the test, and the fact that tense voice quality was used with two different f_0 contours, the test was run twice. In the first run, the target emotion group 'indignation' (with 'VQ only'- tense voice, 'f_0 only'- modal + f_0 'indignation' and 'VQ + f_0' tense voice + f_0 'indignation') was omitted; in the second run the target emotion group 'joy' ('VQ only'- tense voice, 'f_0 only'- modal + f_0 'joy' and 'VQ + f_0' - tense voice + f_0 'joy') was omitted.

Subsequently, a simple contrasts repeated measures ANOVA was conducted in which every stimulus was compared to the modal (neutral f_0), as well as with every other stimulus. The within-subjects factor was Stimulus, and the between-subjects factor was Language.

Mauchley's test of sphericity routinely conducted by the SPSS for the repeated measures ANOVA indicated that the data did not meet the assumption of sphericity. Therefore, the Greenhouse-Geisser corrections were applied to the degrees of freedom in the analysis. To adjust for multiple comparisons, a Bonferroni correction of the alpha value was used in the post-hoc tests. The detailed output for the ANOVA is documented fully in Appendices 2 - 3. The output of the 'group' ANOVA is given in Appendix 2. The results of the simple contrasts ANOVA is given in Appendix 3. The statistical measures of raters' agreement are given in Appendices 4 -6.

It should be noted that statistical results for the group tests are likely to show cross-language differences as an interaction effect. Note that for a simple effect of language to emerge, there would have to be very large differences affecting a number of stimuli in each group. It is important to bear in mind that for the signalling of a particular affect, most of the stimuli may not be all that relevant. Thus, if there is a cross-language effect, it is most

likely to show up in terms of one or two stimuli which are deemed to carry affective colouring in one or more languages. As the 'group' ANOVA pools stimuli in each group such cross-language differences might not emerge as statistically significant in the separate ANOVA test for between-subjects effect (Language). Most cross-language differences will show up as an interaction of the between-subjects effect (language) and the other two ('stimulus type' and 'target emotion group').

4.2 Apologetic-Indignant

4.2.1 The statistical analyses

For the first part of the 'group' ANOVA, (including the 'joy', and excluding the 'indignation' 'target emotion groups') the effects of the following two way interactions were statistically significant: Target emotion group x Language, $F(5.9, 155.8) = 4.9$, $p < 0.001$, partial $\eta^2 = 0.16$; Stimulus type x Language, $F(1.3, 104.7) = 5.7$, $p < 0.001$, partial $\eta^2 = 0.18$, and Target emotion group x Stimulus, $F(3.6, 280.2) = 110.74$, $p < 0.001$, partial $\eta^2 = 0.59$. The three-way interaction (Target emotion group x Stimulus type x Language) was not statistically significant, $F(10.8, 280.2) = 1.56$, $p = 0.11$, partial $\eta^2 = 0.06$.

When the design included the group 'indignation' the three-way interaction (Target emotion group x Stimulus type x Language) was found to be statistically significant: $F(10.7, 277) = 2.01$, $p = 0.03$, partial $\eta^2 = 0.07$. The test of between-subjects effect (Language) was not significant: $F(3, 78) = 0.47$, $p = 0.71$, partial $\eta^2 = 0.02$.

The results are plotted in Figure 4.1. The presence of interaction effects suggests that the difference in affective ratings cannot be explained only as the individual influences of either the stimulus type, the 'target emotion group' or the language. Rather, it is influenced by a particular combination of the levels of these independent factors, with the corresponding consequences for the score of the dependent variable (ratings). The detailed analysis of these factors will be given below, where the results for each of the affects in every paired test are presented.

Analysis of simple contrasts in which modal was compared to all other stimuli showed significant Stimulus x Language interaction $F(8.9, 231.5) = 3.39$, $p = 0.01$, partial $\eta^2 = 0.12$. Between-subjects effect of the language was not significant $F(3, 78) = 2.11$, $p = 0.11$, partial $\eta^2 = 0.08$.

The raters' agreement was as follows:

Hiberno-English:	ICC (r) = 0.60, ICC (R) = 0.97.
Russian:	ICC (r) = 0.78, ICC (R) = 0.99.
Spanish:	ICC (r) = 0.40, ICC (R) = 0.93.
Japanese:	ICC (r) = 0.45, ICC (R) = 0.95.
All:	ICC (r) = 0.53, ICC (R) = 0.99.

The raters' agreement was therefore substantial for Hiberno-English and Russian listeners, and moderate for Spanish and Japanese listeners. The overall agreement was

moderate, suggesting that the raters from different language groups assigned similar values to the individual stimuli.

4.2.2 Apologetic

While there are no (to our knowledge) data in the literature on the vocal signalling of the *apologetic* affect, it could be hypothesised that its expression will be related to a certain extent to the expressions of politeness and guilt. Relatively low vocal effort (and consequently non-tense voice) and low pitch variation could be expected. In Japanese, vocal expression of politeness has been associated with increased breathiness (Campbell, 2004a; Ito, 2005). Ishi, Ishiguro and Hagita (2008) reported increase whisperiness/breathiness in the expression of embarrassment in Japanese. Ohala's frequency code (Ohala, 1994, 1996) would further suggest the use of high-pitched voice in the expressions of *apologetic*.

The 'VQ only' stimuli

Whispery and lax-creaky voice yield the highest ratings for *apologetic*, with evidence for language specific preferences. Whispery voice is associated with this affect by Russian and Hiberno-English listeners, with Spanish and Japanese yielding subthreshold ratings. Lax-creaky voice is favoured by Russian and Spanish listeners, while Hiberno-English listeners give subthreshold ratings, and Japanese ones show no affective colouring for this quality.

For all languages, the ratings for whispery and breathy show, relative to modal, a considerable shift towards *apologetic*, even if the ratings are not very high. This observation also holds for lax-creaky for all languages other than Japanese.

Voice quality does get associated with this affect. Although there is no clear one-to-one voice to affect mapping, the three stimuli that get associated with *apologetic* have breathiness in common (see description of the stimuli in Chapter 3). The findings do not, however, strongly support the literature claims for the Japanese: whispery, breathy or lax-creaky do not achieve high ratings here for this group of listeners. This does not mean that voice quality is entirely ineffective for Japanese: note that relative to modal, whispery shifts the ratings towards *apologetic*, although the effect is still subthreshold. Nonetheless, Japanese stand out relative to the other languages in that apologetic is not strongly cued by any of the present 'VQ only' stimuli.

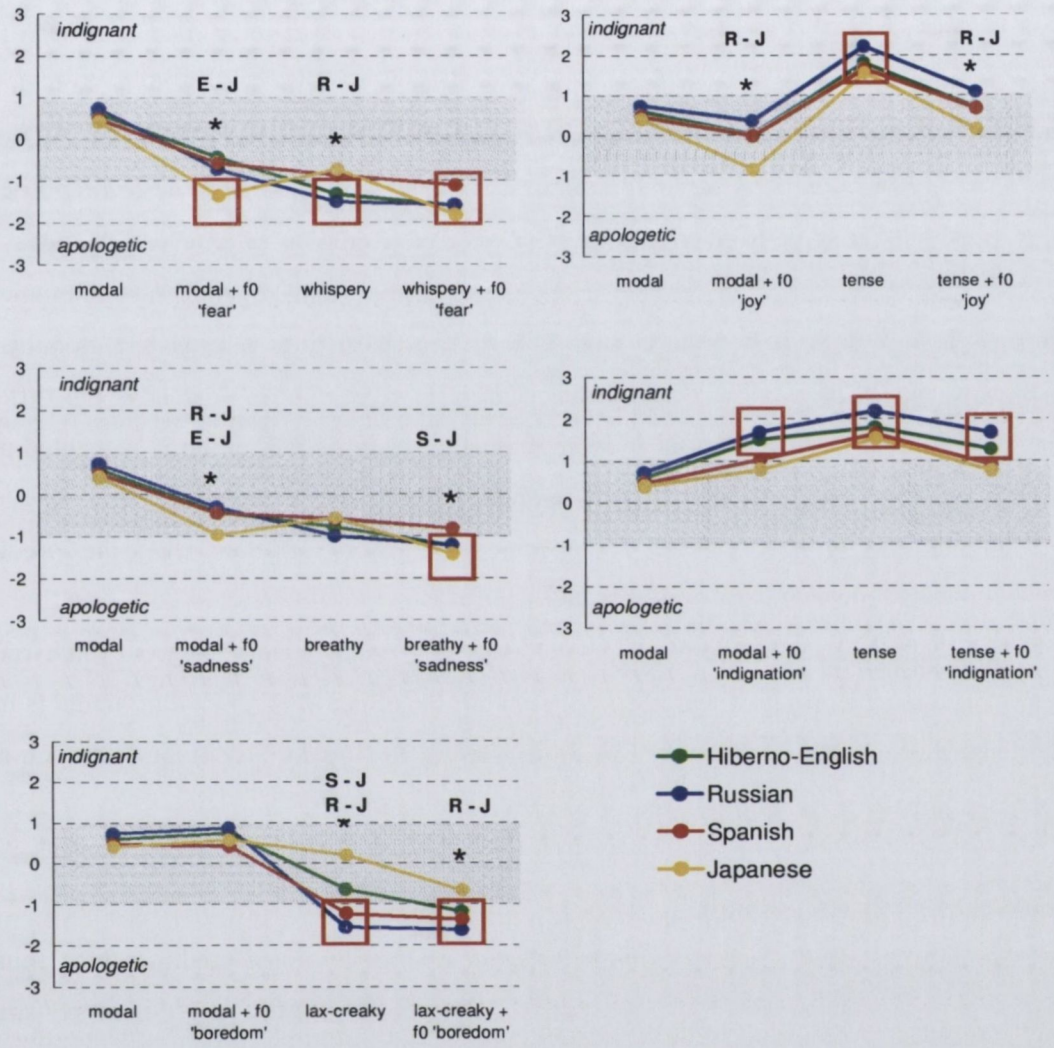


Figure 4.1. *Affective rating for each of the stimuli within the 'target emotion groups'. Asterisks show significant differences among the language groups at $p < 0.05$.*

The 'f_0 only' stimuli

The group of Japanese listeners stands out as different in attaching the importance of f_0 in cueing *apologetic*. This can be seen in the response (relative to modal) of modal + f_0 'fear', modal+ f_0 'sadness' and even modal + f_0 'joy'. For Japanese, having a very high pitch appears to be more efficient in cuing *apologetic* than any (of the present) stimuli from the 'VQ only' series. This supports the literature: high pitch is associated with submissiveness and politeness (Ofuka et al., 2000) as well as with such positive characteristics as cuteness, kindness and politeness (while low pitch in female speakers suggested such negative characteristics as stubbornness, selfishness and strength (Ide & Yoshida, 1999). It is not just the global pitch, but also aspects of the contour that influence the raters judgment. Note how the different dynamics of modal + f_0 'indignation' (a more dramatic and rapid fall on the 2nd accented peak) counter the *apologetic* affective interpretation decisively.

To sum up, simply raising f_0 appears to be more efficient than the predicted breathiness/whisperiness for the Japanese subjects.

The other languages show the same trend to a much smaller degree: pitch raising generally changes the ratings towards *apologetic*) but the effect of f_0 relative to modal is much smaller, and the voice quality cues (of breathy voice or lax creaky voice) appear much more important to the affect signalling in the three European languages.

The combined 'VQ + f_0' stimuli

Combining raised f_0 contours with whispery or breathy voice qualities does enhance affective ratings considerably for the Japanese. For these subjects, the combined stimuli whispery + f_0 'fear' and breathy + f_0 'sadness' give the highest ratings, and the difference relative to breathy and whispery voice is statistically significant.

In the other languages, the effect of adding non-neutral f_0 contours to these voice qualities (whispery and lax-creaky) does not enhance the ratings significantly.

Cross-language similarities/differences

a) Japanese differed from the European languages in that *apologetic* was not strongly signalled by either lax-creaky or whispery voice.

b) Japanese differed from the European languages in that the elevated pitch contour f_0 'fear' does elicit this affect.

c) Among the European languages, there were language-specific preferences in the choice of lax-creaky vs. whispery/breathy voice.

d) Combining the elevated f_0 'fear' with whispery voice did enhance the percept for the Japanese subjects, but not for the European.

Apologetic: summary

- European languages: effectively cued by either whispery or lax-creaky voice, with cross-language preferences.
- European languages: none of the f_0 contours on its own cues this affect.
- No one-to-one voice to affect mapping.
- Japanese: the raised f_0 'fear' contour is effective, on its own, but particularly when combined with whispery voice.
- The present 'VQ only' stimuli do not efficiently signal apologetic for the Japanese.

4.2.3 Indignant

Indignation is a form of anger, so in the expression of indignation acoustic manifestations of anger can be expected: tense or harsh voice quality as well as quite a substantial variability and range of f_0, high mean intensity (increased vocal effort) (Burkhardt & Sendlmeier, 2000; Laukkanen, Vilkman, Alku, & Oksanen, 1996; Scherer, 2003), see also Table 2.4 from Chapter 2 for summary of the literature findings.

The 'VQ only' stimuli

Tense voice yields the highest ratings for all languages, significantly higher than those of modal. The different languages appear uniform in selecting this stimulus, although some minor (not significant statistically) differences in the relative strength of affective ratings can be observed. This supports traditional descriptions found in the literature. Note, however, that although anger/indignation is often associated in the literature with an increased f_0 range, no f_0 variation was required for the listeners to associate this voice with *indignation*.

The 'f_0 only' stimuli

A similar trend is observed for all languages: f_0 'indignation' (modal + f_0 'indignation') communicates *indignant*, in particular for Hiberno-English and Russian listeners (the ratings are subthreshold for Spanish and Japanese). However, the difference between the ratings of modal and modal + f_0 'indignation' does not reach statistical significance.

The combined 'VQ + f_0' stimuli

The addition of f_0 'indignation' to tense voice does not enhance the impression of *indignant* for all languages. Quite the contrary, the combination appears counterproductive as the rating decreases relative to tense for all these languages, although the difference does not reach statistical significance. (Note also the similar and even more dramatic effect of combining tense voice with f_0 'joy'.) The addition of the extreme f_0 contour to either the modal or the tense voice quality also decreases the degree of certainty in the affective responses (as seen from the high interquartile range for Spanish, Russian and Japanese, Appendix 6 and box plots in Appendix 1. This was somewhat a surprise finding, which goes against our initial expectation that adding extreme f_0 perturbations might enhance the perception of the strong emotions. It may be that the expression of indignation does not require simultaneous presence of both voice and f_0 cues.

Cross-language similarities/differences

The tendency is similar across all language groups in that tense voice quality on its own seems to be more effective than the f_0 contour 'indignation'. Surprisingly, no affective enhancement is achieved by combining the two. For all the language groups tested, the combination of tense + f_0 'indignation' was less effective than tense voice with the neutral f_0 contour.

Indignant: summary

- One voice to cue indignation for all languages – tense.
- f_0 'indignation' on its own does signal this affect, but does not achieve what tense voice on its own can.
- Combining tense and f_0 'indignation' was ineffective and did not enhance the affective ratings.
- Similar trends were observed across all language groups.

4.3 Bored-Interested

4.3.1 The statistical analyses

The results of statistical analysis showed the following. For the first part of the 'group' ANOVA (excluding the emotion target group 'indignation'), the three-way interaction (Target emotion group x Stimulus type x Language) was statistically significant: $F(7.3, 191) = 2.11$, $p = 0.04$, partial $\eta^2 = 0.08$. The test of between-subjects effect (Language) was not statistically significant: $F(3, 78) = 0.77$, $p = 0.52$, partial $\eta^2 = 0.03$.

When the design included the group 'indignation' the three-way interaction was also statistically significant: $F(8.9, 231.2) = 2.23$, $p = 0.02$, partial $\eta^2 = 0.08$. The test of between-subjects effect (Language) was not statistically significant: $F(3, 78) = 0.86$, $p = 0.47$, partial $\eta^2 = 0.03$.

The results are plotted in Figure 4.2. As in the previous subtest, the presence of three-way interaction between Emotion x Stimulus x Language suggests that a particular combination of the levels of these independent factors influenced the affective ratings. The detailed analysis of the simple effects will be given below, separately for each of the affects.

Analysis of simple contrasts in which modal was compared to all other stimuli showed significant Stimulus x Language interaction, $F(10.9, 285.3) = 3.06$, $p = 0.01$, partial $\eta^2 = 0.1$. Between-subjects effect of language was not significant $F(3, 78) = 0.97$, $p = 0.41$, partial $\eta^2 = 0.04$.

The raters' agreement was as follows:

Hiberno-English:	ICC (r) = 0.52, ICC (R) = 0.96.
Russian:	ICC (r) = 0.67, ICC (R) = 0.98.
Spanish:	ICC (r) = 0.55, ICC (R) = 0.96.
Japanese:	ICC (r) = 0.73, ICC (R) = 0.98.
All:	ICC (r) = 0.60, ICC (R) = 0.99.

The raters agreement was substantial for the groups of Russian and Japanese subjects, and moderate for the Hiberno-English and Spanish subjects. The overall agreement was substantial. This is again in keeping with the results of ANOVA that showed no significant effect of the between-subjects factor Language. The results suggest that the listeners rated the stimuli in this subtest in a rather consistent and uniform way and that no cross-cultural differences in ratings are likely to be observed.

4.3.2 Bored

Boredom has traditionally been associated with creaky voice (e.g., Laver, 1980). Previous perception tests on Hiberno-English subjects (Gobl & Ní Chasaide, 2003b) have suggested that a lax creaky voice provides is particularly potent cue. Other features associated in the literature with boredom are a low mean, floor and range of f_0, flat monotonous f_0 contour, rather low mean intensity, slow speech rate and low articulation precision (refs here) See the summary in Table 2.4, Chapter 2.

The 'VQ only' stimuli

As expected and in keeping with the literature claims, lax-creaky voice yields the highest ratings for *bored*, for all languages. *Bored* is also associated (less strongly) with whispery voice and (even more weakly) with breathy voice. There are some cross-language differences. Whispery does not signal boredom for Hiberno-English listeners, and breathy yields near threshold ratings for Japanese. The ratings of these voice qualities are on the whole significantly higher relative to modal.

Although more than one voice quality gets associated with boredom, lax-creaky appears to be the most effective for all these languages. On the other hand, the difference in ratings for lax-creaky, breathy and whispery in most cases does not reach statistical significance.

The 'f_0 only' stimuli

As there was only a very small difference between the f_0 contours 'boredom' and 'neutral' (see Figure 3.3 in Chapter 3), it is not surprising that the ratings of modal and modal + f_0 'boredom' show no significant difference. None of the two stimuli was associated with boredom by the language groups tested. Flat f_0 and low f_0 level, although essential in signalling boredom, are clearly not sufficient.

The combined 'VQ + f_0' stimuli

No significant difference in ratings was found between lax-creaky on its own and lax-creaky combined with f_0 'boredom' for all languages. Note that lax-creaky from the 'VQ only' series has intrinsic f_0 about 30 Hz lower relative to modal (neutral f_0) (see Figure 3.3 in Chapter 3), and as a consequence, its f_0 is also lower than that of the combined stimulus. This difference came largely unnoticed by the listeners, although there is a clear trend for the

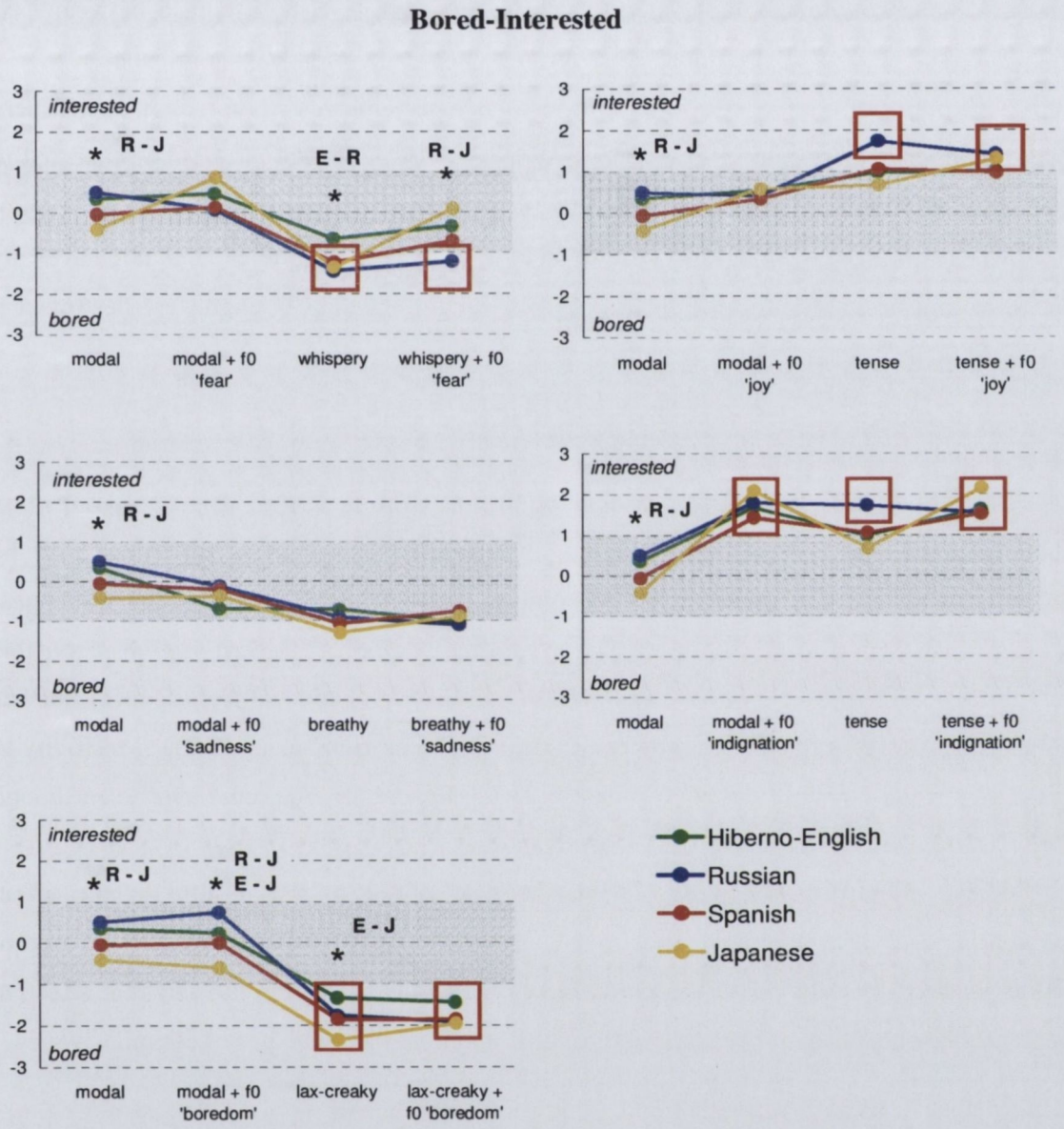


Figure 4.2. *Affective rating for each of the stimuli within the 'target emotion groups'. Asterisks show significant differences among language groups at $p < 0.05$.*

Japanese listeners to favour the lower f_0 of lax-creaky voice. The difference between lax-creaky and lax-creaky + f_0 'boredom', however, does not reach statistical significance.

One potentially interesting cross-cultural difference emerges when ratings of whispery voice and whispery + f_0 'fear' are compared. Whispery voice was not intended or expected to signal boredom, but it did for all language groups except for Hiberno-English. The addition of f_0 contour 'fear' (high f_0 level and considerably greater dynamic range) did not have any significant impact on the impression of boredom for the Russian listeners, yet destroyed it completely (significantly) for the Japanese.

Cross-language similarities/differences

The trends in voice to affect association are very similar for all language groups: lax creaky voice yielded highest ratings for all. A possible cross-language difference may be indicated in the perception of the combined stimulus whispery + f_0 'fear'. The Japanese appear more sensitive (again) to f_0 variation as seen from the addition of f_0 'fear' to whispery: this destroys the effect of boredom for the Japanese completely (statistical significance) but has only a minor such effect for other listeners. Perhaps a similar trend is appearing (sensitivity to f_0 variation) in the Japanese ratings for lax-creaky vs. lax-creaky + f_0 'boredom' even though the effect here does not reach statistical significance.

These effects may suggest (as was also suggested in the results for the affect *apologetic above*, that there may be cross-language differences for the Japanese compared to the European group in the relative importance of f_0 and VQ cues and in the way that affective ratings may be enhanced or disenhanced through combining the two types of stimuli.

Bored: summary

- Boredom is cued by lax-creaky voice for all language groups, but also, with cross-language preferences, by whispery and breathy.
- Low level and range of f_0 'boredom' are not enough to signal this affect, so voice quality does appear to be crucial to this affect.
- No difference between the ratings of lax-creaky and lax-creaky + f_0 'boredom'.
- Similar cross-language trends. Note however, that the Japanese appear to be more sensitive to the f_0 cues compared to other languages.

4.3.3 Interested

Although this affective state has not been extensively discussed in the literature, the expectations/predictions for the vocal expression of interest are those opposite to boredom: the absence of creaky component in phonation, possibly modal to tense voice, a raised f_0 level and a more dynamic f_0 contour. One might surmise that the expression and communication of interest might like joy rely somewhat on facial and postural expressive cues.

The 'VQ only' stimuli

For all these languages the ratings of tense voice show a clear shift towards *interested* relative to modal. The ratings, however, are high for the Russian listeners only, but near- to subthreshold for the other language groups, being lowest for Japanese. [Note also that for the Japanese the interrater agreement/percent agreement for tense voice was rather low (see Appendix 6)] Against expectations, modal voice does not achieve a high rating here. Breathy/whispery voice in Japanese has been reported to express polite interest (Ishi et al. 2008), but no such association has been found for the stimuli from this set.

The 'f_0 only' stimuli

There was no major expectation as to whether the f_0 contour would contribute to affect signalling here, as we did not have a special 'interested' contour. However, one might broadly expect those contours which have a raised f_0 level with a large dynamic range might cue for interest. These would be the f_0 contours modal + f_0 'indignation', modal + f_0 'joy' and modal + f_0 'fear'.

Looking at responses for those stimuli, we do see a rather uniform trend across languages. The most highly rated for *interested* was modal + f_0 'indignation', and the effect is significant for all four languages. Interestingly, strong cueing of interest did not emerge with f_0 'joy' or f_0 'fear'. This difference in reaction possibly has to do with the difference in the range and dynamics of the contours. As evident from the comparison of the contours (see Figure 3.3, Chapter 3), f_0 'indignation' is of substantially higher level and has a more rapid fall in the second peak than f_0 'joy' or 'fear'.

Although all languages show clear agreement here, the magnitude of the difference is the largest for the Japanese listeners. These listeners stand out (again) as showing higher sensitivity/attaching relatively more importance to f_0 cues. This is also evident in the

'indignation' group where the difference between tense voice with and without f_0 'indignation' reaches statistical significance for the Japanese listeners only.

The combined 'VQ + f_0' stimuli

Combining tense voice and dynamic f_0 ('indignation' and 'joy') showed the following language-specific trends. For Russian the ratings for tense and tense + f_0 'indignation' (as well as modal + f_0 'indignation') are virtually the same. For other languages, and in particular for the Japanese, there is a clear trend towards lowering in the ratings if the f_0 is neutral, although, as stated above, the difference reaches statistical significance for the Japanese only. It can be concluded that for the languages tested here other than Russian, the high, dynamic f_0 appears to be the major cue while the Russians appear to use both cues to the same extent.

When tense voice is combined with f_0 'joy', no significant enhancement of ratings for *interested* can be observed for the Russian, Spanish or Hiberno-English listeners. Although there is a trend towards increase in ratings for the Japanese, it does not reach statistical significance.

Cross-language similarities/differences

All languages appear to rate stimuli with very high dynamic f_0 ('indignation') as expressing interest. A greater sensitivity to the presence/absence of f_0 cues is showed by the Japanese listeners as evident from the ratings of tense + f_0 'indignation' relative to tense (ratings for *interested* increase when f_0 is high and dynamic). Pointers that the Japanese subjects may pay greater attention to f_0 cues were also found in the results for *apologetic* and *bored.* Note that Russian listeners rated these 'VQ only' and 'VQ+ f_0' stimuli in exactly the same way.

Russian showed bias in rating the tense voice for *interested* higher than other groups of listeners. For this group of listeners, voice quality and f_0 cues appear equally important.

Interested: summary

- *Interested* is not cued well by voice quality on its own; tense voice yielded high ratings for Russian listeners only.
- One specific f_0 contour 'indignation' (not just any high, dynamic contour available) signals *interested* for all languages.

- Combining tense voice and f_0 'indignation' does enhance the ratings of tense, but clearly f_0 acts as a primary cue: the combined stimulus tense voice + f_0 'indignation' does not achieve higher rating relative to modal + f_0 'indignation'.
- Language differences are manifested in the relative sensitivity to f_0 and voice quality cues: tense on its own can cue *interested* for Russians but not for the other languages; Japanese listeners appear most sensitive to the presence of f_0 cues.

4.4 *Intimate-Formal*

4.4.1 The statistical analyses

The results of statistical analysis can be summarised as follows. For the first part of the group ANOVA, the three-way interaction (Target emotion group x Stimulus type x Language) was statistically significant, $F(9, 234.6) = 14.14$, $p < 0.001$, partial $\eta^2 = 0.35$. The test of between-subjects effect (Language) was statistically significant: $F(3, 78) = 10.4$, $p < 0.001$, partial $\eta^2 = 0.29$.

When the design included the group 'indignation' the three-way interaction was also found statistically significant: $F(8.9, 230.8) = 5.1$, $p < 0.001$, partial $\eta^2 = 0.16$. The test of between-subjects effect was statistically significant: $F(3, 78) = 3.47$, $p = 0.02$, partial $\eta^2 = 0.12$.

These results suggest quite prominent cross-language differences in the way affective ratings were assigned to particular stimuli.

The results are plotted in Figure 4.3. The presence of a significant three-way interaction effect suggests again that the difference in affective ratings depends on what 'target emotion' group the stimulus belongs to – that is, the type of voice quality and the kind of f_0 contour used. These factors will be analysed below, where each of the affects *intimate* and *formal* are discussed along with the associated stimuli.

The analysis of simple contrasts in which modal was compared to all other stimuli showed a significant Stimulus x Language interaction $F(8.7, 227.4) = 16.9$, $p < 0.001$, partial $\eta^2 = 0.4$. The between-subjects effect of language was not significant $F(3, 78) = 0.96$, $p = 0.41$, partial $\eta^2 = 0.04$.

The raters' agreement was as follows:

Hiberno-English:	ICC (r) = 0.52, ICC (R) = 0.96.
Russian:	ICC (r) = 0.78, ICC (R) = 0.99.
Spanish:	ICC (r) = 0.30, ICC (R) = 0.90.
Japanese:	ICC (r) = 0.27, ICC (R) = 0.89.
All:	ICC (r) = 0.19, ICC (R) = 0.95.

The raters' agreement varied to a larger extent for this test compared to the previous two. While the raters' agreement was substantial for Russian listeners and moderate for Hiberno-English listeners, it was poor for Spanish and Japanese listeners. The overall agreement was also poor. The results suggest that there was considerable disagreement in the

way certain stimuli were associated with this affect by members of the different language groups. Therefore, in this test prominent cross-language differences are expected to emerge.

4.4.2 Intimate

Intimacy is quite a complex affective phenomenon, it implies closeness in terms of social distance, confidentiality and familiarity (Seki et al., 2002), and has been reported to be associated with breathy or whispery voice (Laver 1980). In perception studies (Gobl et al., 2002; Gobl & Ní Chasaide, 2003b), intimacy was found to be strongly associated with lax-creaky voice for Hiberno-English subjects, and to a lesser degree with breathy and whispery.

Dynamic f_0 is not intuitively related to intimacy, at least in adult speech, but see evidence for increased pitch and pitch range in child-directed speech as in Batliner, Schuller, Schaeffler and Steidl (2008).

The 'VQ only' stimuli

Whispery and lax-creaky voices were associated with *intimate,* but only by Hiberno-English and Russian listeners. Of these two qualities, lax-creaky yielded somewhat (but not significantly) higher ratings. For Russian only, breathy voice yielded near-threshold mean ratings. For these languages, whispery, breathy and lax-creaky voices show, relative to modal, a sizeable shift towards *intimate*.

There was quite a prominent cross-cultural difference, as none of these voice qualities yielded strong ratings for *intimate* for the Spanish and Japanese listeners. Lax-creaky yields a mean value around zero for Spanish and Japanese; the listeners appear to assign very disparate ratings to the stimulus and the spread of responses is very broad (as shown by the large interquartile range in Appendix 1 or the percent agreement data in Appendix 4). Note also that *intimate* does get associated by these groups of listeners with the modal + f_0 'indignation' and the tense + f_0 'indignation' stimuli. This goes against expectations, but listeners are quite consistent in their judgments and the percent agreement is high.

The finding that lax-creaky and whispery voice signal intimacy for Russian and Hiberno-English was in keeping with our initial expectations in this experiment, given the results of some previous experiments, e.g., Gobl and Ní Chasaide (2003b), Gobl et al. (2002). Note that both these voice qualities have a breathy/whispery quality in common, and that

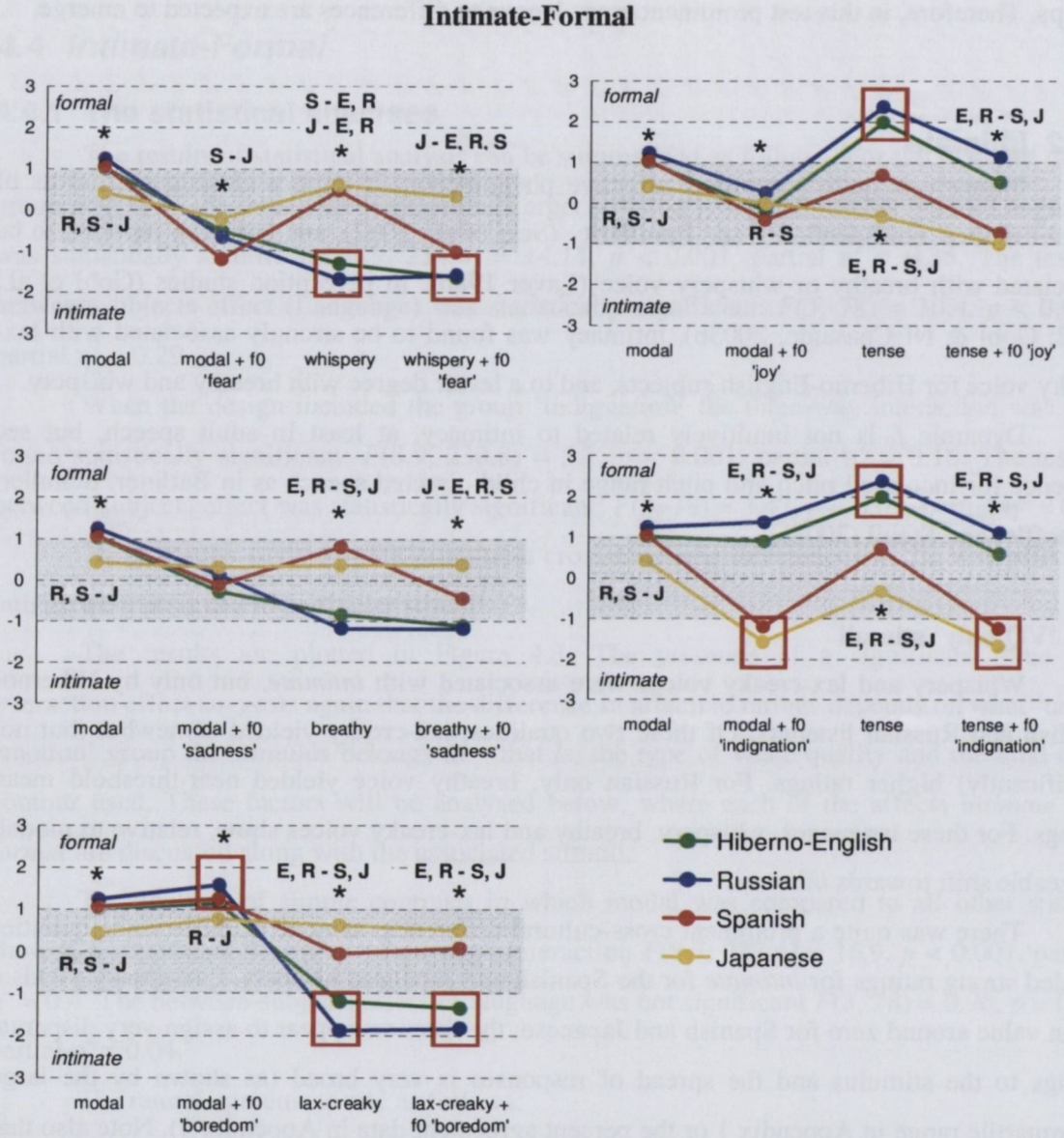


Figure 4.3. *Affective rating for each of the stimuli within the 'target emotion groups'. Asterisks show significant differences among language groups at $p < 0.05$.*

traditionally in the phonetics literature, breathy or whispery voice have been associated with intimacy.

What is striking here is the fact that these voice qualities do not appear to be universally associated with intimacy.

The 'f_0 only' stimuli

The data shows that the f_0 contour 'indignation' signals *intimate* for the two very different languages, Spanish and Japanese. The cross-language difference here is striking as the same stimulus is associated with a relatively high degree of formality for Russian and Hiberno-English subjects. It seems clear that the interpretation of this very high and dynamically varied contour was very different for these two groups. The difference between them is statistically significant.

If we look at the responses to modal + f_0 'fear' we can make some inferences about what might be going on here. Note that this latter stimulus has also a very high f_0 level, but it does not have the extensive dynamic fall on the final accent. In this case, responses for the Spanish subjects are still rather high on intimacy, but for the Japanese the effect is gone. From this we could tentatively conclude that the high level pitch is crucial for the Spanish, whereas for the Japanese, the high pitch may need to be linked to the extreme dynamics exemplified in the f_0 indignation contour. Finally, note that for f_0 contour 'joy' which also has a quite elevated f_0 level, ratings for both these languages are close to zero.

The dramatic extent to which the f_0 contour 'indignation' shapes the affective response for Spanish and Japanese is highlighted by the extent of the shift from modal (with neutral f_0).

The combined 'VQ + f_0' stimuli

For Russian and Hiberno-English subjects, lax-creaky and whispery voice were strongly associated with *intimate* and breathy voice to a lesser degree. These same stimuli (lax-creaky and whispery voice) did not signal *intimate* for the Japanese subjects.

The addition of f_0 'boredom', f_0'fear' or f_0 'sadness' had little impact on the ratings for *intimate* relative to the corresponding 'VQ only' (lax-creaky, whispery, breathy) stimuli. Given that quite a number of different f_0 contours were available to the listener, one would tend to conclude that for these two languages voice quality is overwhelmingly important and f_0 factors may be largely irrelevant.

For the other group, the Spanish and Japanese, the trend is very different. Here the very high and dynamically varying f_0 contour 'indignation' in combination with tense (or modal) voice appears to best signal *intimate*. Note, however, that the ratings for the combined stimulus tense + f_0 'indignation' is largely the same as for the f_0 'indignation' stimulus (which has modal voice). From this we would tend to conclude that the f_0 cue appears to be dominant. Note also the higher percent agreement for modal + f_0 'indignation' and tense + f_0 'indignation' for Japanese and Spanish subjects compared to tense (Appendix 4), showing a higher degree of consistency and probably confidence in ratings.

These same stimuli (modal + f_0 'indignation' and tense + f_0 'indignation') did not signal *intimate* for the other subjects, but quite the opposite in that they were associated with *formal*.

Cross-language similarities/differences

It is quite unexpected that tense and modal voice combined with f_0 'indignation' are associated with *intimate* for Spanish and Japanese (but with *formal* for Hiberno-English and Russian). Hiberno-English and Russian raters favoured lax-creaky and whispery voice highly for *intimate*, but these qualities were not at all so perceived by the Japanese subjects.

On the face of it, this suggests a very major cross-language difference in the mapping of voice quality to affect for these two groups of languages. But before we conclude on this, it is necessary to consider whether there might be other factors contributing to these results.

One possible factor could be that the affective terms used in the different languages are not (at least fully) equivalent. As mentioned above, intimacy is a complex construct and the possibility exists that the listeners attached somewhat different meanings to the anchor labels when judging the stimuli.

The semantic equivalence of the labels translated from English into the three languages has been discussed in Chapter 3, along with a post-hoc test which was carried out to examine this issue for Japanese. As discussed in that chapter, the nearest equivalent terms, favoured by our translators (and by the Japanese subjects in the later test) appeared to have a somewhat narrower semantic field: while it entailed 'close, friendly, familiar', it did not include the possible sexual connotations of the English term. In Spanish, the choice of label that was deemed the most equivalent by the translators did not include the sexual overtones either. Note that in Russian, the term used does allow for the broader interpretation as in English (including sexual overtones).

It is quite possible therefore that this difference in semantic fields may have contributed to the dramatically different results shown here for intimacy. Intuitively it seems unlikely that the semantic differences we are discussing here could in themselves account for the very large differences in the voice qualities that were highly rated for intimacy: Spanish and Japanese listeners systematically associate *intimate* (= 'friendly, familiar, close') primarily with dynamic f_0 combined with tense/modal voice quality (stimuli which conjure up formality for the Hiberno-English and Russian listeners). Lax-creaky or whispery voice quality are associated with *intimate* for Hiberno-English and Russian listeners and not at all for the Japanese and Spanish subjects.

The fact that lax-creaky and lax-creaky + f_0 'boredom' yield zero affective ratings for the Spanish and Japanese raises another possibility. To my ear, the neutral voice of Spanish is very different from that of English or Russian. If the neutral voice in Language A is different from Language B (and intuitively this seems to be the case), and if the neutral voice veers towards a quality which is affectively loaded in Language B, then it stands to reason this affect is going to be differently signalled in Language B.

These possibilities cannot be resolved within the confines of the present study, but they do point to some interesting directions for future development of this research. It would be interesting, for example, to design a future experiment where the listeners are presented with the stimuli as in the present study and rate them first in broad 'umbrella' terms of emotion 'families', as in Baron-Cohen (2007), (e.g., 'familiar') and then are asked to specify the association in more narrow terms (e.g., 'friendly', 'close', 'intimate' etc.).

Intimate: summary

- Very clear cross-language difference, Russian and Hiberno-English vs. Spanish and Japanese, both in terms of the relative weight attached to f_0 vs. VQ cues and in the selection of voice qualities.
- Lax-creaky and whispery signal *intimate* for Russian and Hiberno-English, but not for Spanish and Japanese.
- Tense + f_0 'indignation' signal *intimate (familiar)* for Spanish and Japanese, but not for Hiberno-English and Russian.
- Unclear whether some differences in the semantic scope of the translated labels might contribute to this result.

- Unclear whether different neutral baseline voice qualities in the different languages might partially explain the differences in results.

4.4.3 Formal

Formality is generally described as being conveyed by a decreased vocal expressiveness, so the use of the modal voice, low f_0 range and decreased f_0 variation are to be expected. An increased vocal effort and articulation precision are further expected along with some postural adjustments (erect and tense posture) and increased interpersonal distance (Burgoon & Hoobler, 2002). In Japanese the expression of politeness (essential in the formal contexts) can be expressed by the use of breathy voice (Ito, 2005).

The 'VQ only' stimuli

There are some cross language differences here. For Hiberno-English and Russian, tense voice is strongly suggestive of formality, but this is not at all the case for Spanish and Japanese. Modal voice (with neutral f_0) yields a weak, near-threshold effect in all but Japanese. For Hiberno-English and Russian, the ratings for tense voice are significantly higher than the modal ratings.

Tense voice yielded very variable responses from the Japanese and Spanish, which essentially brought the mean value close to zero (see Appendix 1 for interquartile range and percent agreement Appendix 4 or scatter plots in Appendix 6).

It has been shown experimentally that the expression of politeness (essential in the formal contexts) can be expressed by the use of breathy voice (Ito, 2005). However, breathy voice was not selected as signalling formality here, nor indeed other stimuli that included settings for breathiness, i.e. lax-creaky or lax-creaky + f_0 'boredom'.

The 'f_0 only' stimuli

As predicted from the literature (Burgoon & Hoobler, 2002), large f_0 excursions are not associated with formality. While modal voice with the neural f_0 contour (weakly) signals formality for the three European languages, the addition of the dynamic f_0 contours ('fear', 'sadness', 'joy') lowers the ratings dramatically.

Surprisingly, the f_0 contour 'indignation' yields different results, and the trend is different for the Spanish listeners on the one hand and the Russian and Hiberno-English ones

on the other. The addition of f_0 'indignation' to the modal voice quality shifts the ratings of the Spanish listeners from *formal* to *intimate/familiar*, and the difference is statistically significant. No such effect is seen for the Hiberno-English or Russian listeners for whom this combination is still acceptable as *formal*. Note however that both these stimuli yield considerably lower ratings than tense voice.

The combined 'VQ + f_0' stimuli

The addition of a high, dynamically varying f_0 to tense voice (as in f_0 'indignation' and f_0 'joy') appears to reduce its potency as an expression of formality (for Russian and Hiberno-English). None of the combined stimuli of this test signalled formality for the Spanish or Japanese subjects.

Cross-language similarities/differences

Prominent cross-language differences can be observed. Most conspicuously, none of the stimuli cued *formal* for the Japanese. Furthermore, while tense voice signals *formal* efficiently for Hiberno-English and Russian (significantly more so relative to modal) no such association has been found for the Spanish.

Different sensitivity to f_0 'indignation' can be further observed. For Spanish, modal voice has weak association with *formal*, but dynamic f_0 'indignation' shifts the ratings to *intimate/familiar*. For Russian and, to a lesser extent, Hiberno-English listeners, this combination signals formality.

The striking finding that none of the 'VQ only' stimuli in the present set signalled *formal* for the Japanese, raises some interesting questions. As discussed in Chapter 3, following these results, a post –hoc test was carried out to examine the possibility that non equivalence in the terms used might have contributed to differences such as these. It was concluded that the term used (かしこまった 'formal, psychologically distant') indeed corresponded to the intended meaning in English and that it would not appear to have caused interpretation difficulties to the Japanese subjects. Labelling issues can hardly be invoked to explain the almost total absence of formal affect signalling from our stimulus set here.

Another factor that may be highly relevant here is the fact that the expression of formality in Japanese involves a system of honorifics – *keigo* ('terms of respect') (e.g., Ito, 2005; Ofuka et al., 2000). One could reasonably speculate therefore that voice quality may very well not be used, or if used is of relatively minor importance.

Formal: summary

- Prominent cross-language differences.
- *Formal* is effectively cued by tense voice for Hiberno-English and Russian listeners.
- *Formal* is weakly associated with modal voice for all the languages but Japanese.
- None of the stimuli cued *formal* for the Japanese.
- The addition of dynamic f_0 contours to tense voice lowers the affective ratings.

4.5 Relaxed-Stressed

4.5.1 The statistical analyses

Statistical analysis showed the following results. For the first part of the ANOVA, the three-way interaction (Target emotion group x Stimulus type x Language) was statistically significant, $F(8.4, 217.7) = 5.8$, $p < 0.001$, partial $\eta^2 = 0.18$. The test of between-subjects effect was statistically significant: $F(3, 78) = 6.14$, $p = 0.001$, partial $\eta^2 = 0.19$.

When the design included the group 'indignation' the three-way interaction was also statistically significant: $F(10.6, 274.5) = 4.18$, $p < 0.001$, partial $\eta^2 = 0.14$. The test of between-subjects effect (Language) was statistically significant: $F(3, 78) = 3.43$, $p = 0.021$, partial $\eta^2 = 0.12$.

The results are plotted in Figure 4.4. As suggested by the three-way interaction between Target emotion group x Stimulus type x Language, the affective ratings are influenced by a particular combination of these independent factors. The detailed analysis of the simple effects will be given below, separately for each of the affects in the *relaxed-stressed* subtest.

Analysis of simple contrasts in which modal was compared to all other stimuli revealed significant Stimulus x Language interaction $F(8, 208.5) = 7.42$, $p < 0.001$, partial $\eta^2 = 0.22$. Between-subjects effect of language was not significant $F(3, 78) = 2.03$, $p = 0.12$, partial $\eta^2 = 0.07$.

The raters' agreement was as follows:

Hiberno-English:	ICC (r) = 0.55, ICC (R) = 0.96.
Russian:	ICC (r) = 0.56, ICC (R) = 0.96.
Spanish:	ICC (r) = 0.72, ICC (R) = 0.98.
Japanese:	ICC (r) = -0.03, ICC (R) = -0.04.
All:	ICC (r) = 0.34, ICC (R) = 0.98.

As in the previous test, there was a noticeable difference in raters' agreement across the language groups tested. The raters' agreement was moderate for Hiberno-English and Russian listeners, substantial for the Spanish listeners and very poor for the Japanese listeners. The overall agreement was also poor, pointing at possible prominent differences between the European languages on the one hand and Japanese on the other in the way the stimuli were associated with affect. See, for example, Appendix 6: whereas responses for the three

European languages show considerable clustering (interrater agreement), there is an almost random scatter of values from Japanese subjects in this test.

4.5.2 Relaxed

It could be predicted that the *relaxed* state is associated with overall muscular relaxation and low (by definition) sympathetic arousal, hence a rather lax (breathy) voice quality, and rather unexcited intonation patterns (low level and range of f_0). Lax-creaky voice was found to signal the *relaxed* state in Gobl and Ní Chasaide (2003b). Relaxed breathing and balanced muscle tone, low f_0, and low to medium speech intensity have been further described for this affect in Johnstone, van Reekum and Scherer (2001).

The 'VQ only' stimuli

Lax-creaky was the voice quality most strongly associated with *relaxed* for all the language groups tested except the Japanese. Whispery and breathy voice were also rather effective in signalling *relaxed* in the European languages. Lax-creaky, whispery and breathy voice show, relative to modal, a considerable shift towards *relaxed*. Note however that the difference between lax-creaky, whispery and breathy voice in these three languages does not reach statistical significance. We conclude therefore, for the European languages, that voice quality maps well to the *relaxed* affect, but note as in many other cases, that it is not a one-to-one mapping and that there is no cross-language differences within this group.

It is striking that none of the present stimuli signal *relaxed* for the Japanese subjects. The average rating fluctuates around zero, but as can be seen in Appendix 6, this is a very wide and almost random scatter of responses in the *relaxed-stressed* subtest. This uncertainty in rating is also evident from a very broad interquartile range of responses for all stimuli (Appendix 1) or a very low percent agreement (Appendix 4).

The 'f_0 only' stimuli

There were no expectations that any of the dynamic or high level f_0 contours would signal the *relaxed* affect, and the f_0 contours with high f_0 levels and high dynamic ranges were indeed not associated with relaxed. It was anticipated that a relatively low or flat f_0 contour (best exemplified in our experiment by the modal + f_0 'boredom' stimulus) might play a role in signalling *relaxed*. This was not found to be the case.

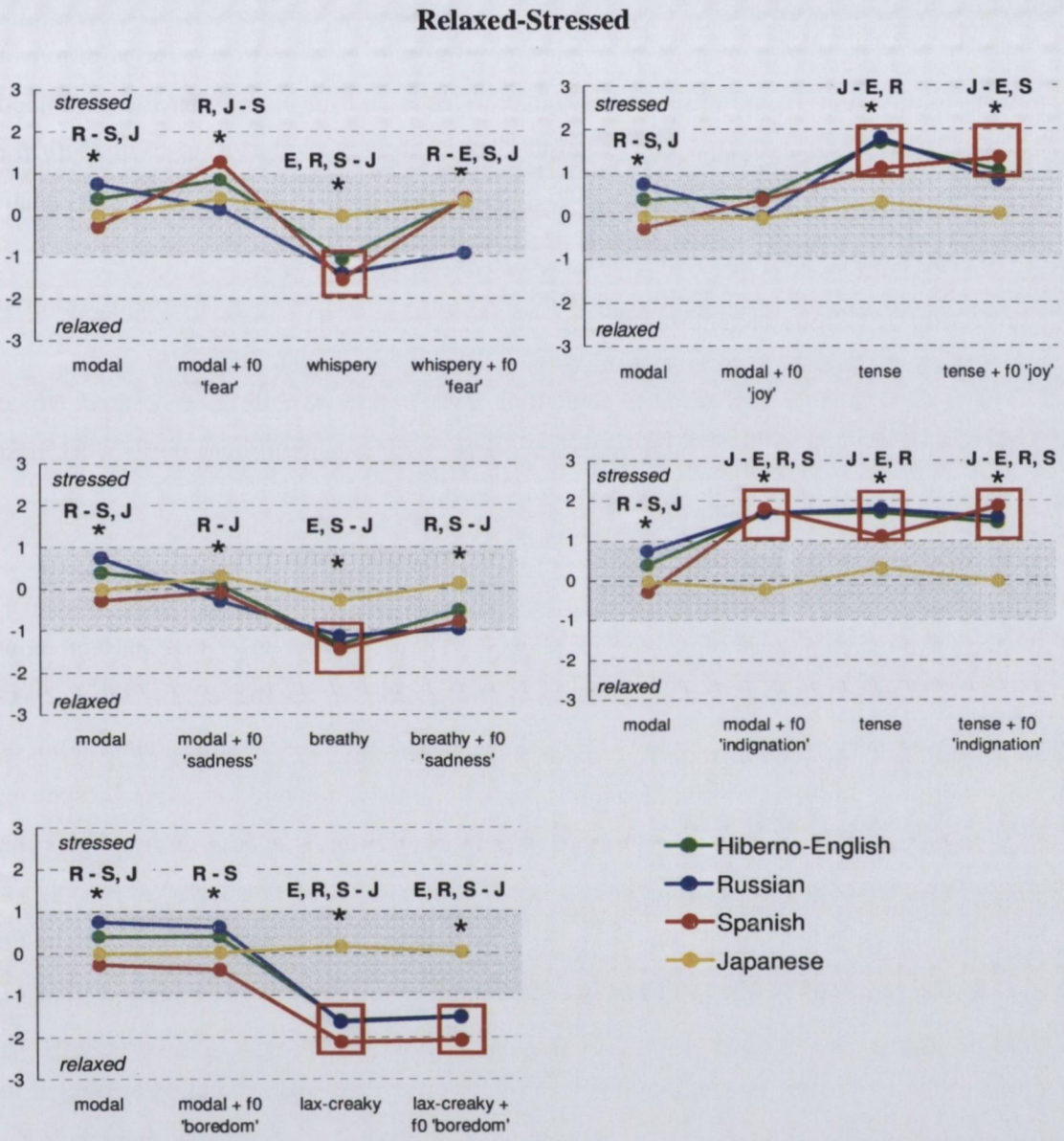


Figure 4.4. *Affective rating for each of the stimuli within the 'target emotion groups'. Asterisks show significant differences among language groups at* $p < 0.05$.

The combined 'VQ + f_0' stimuli

Similarly, no enhancement of affective colouring was expected from the addition of the dynamic f_0 contours to non-modal voice qualities. It is not surprising that the combined stimulus whispery + f_0 'fear' yields significantly lower ratings for *relaxed* than whispery for Hiberno-English and Spanish (though not for Russian). Similar can be observed when comparing the ratings of breathy and breathy + f_0 'sadness'.

No difference is observed for the combined stimulus lax-creaky + f_0 'boredom' relative to the 'VQ only' stimulus lax-creaky (as in other subtests and rather as expected).

The results support our initial expectations that for the signalling of *relaxed*, voice quality is crucial (cf. ratings of modal and non-modal voice qualities), and these work best when f_0 remains relatively neutral (cf. 'VQ only' vs. 'VQ + f_0' stimuli).

Cross-language similarities/differences

There were striking cross-language differences. The results for the *relaxed* affect show similarities both in trends and stimuli choices for the European languages, for whom lax-creaky voice, whispery voice or breathy voice all seem effective. On the other hand, none of the set of 15 stimuli presented to the Japanese listeners in this test elicited a *relaxed* response. Indeed, the Japanese ratings showed enormous variability between subjects, and one gets the impression they were very unsure about how to conduct the ratings for this test.

As mentioned in Chapter 3, when results emerged, one wondered whether the choice of labels or the familiarity with the words might have contributed to this high degree of uncertainty in responses. This led to the post-hoc translation test (Section 3.3.2 of Chapter 3). The results of the latter test suggest that this lack of affective response can not be attributed to the possible difficulties that the raters might have interpreting the translated affective labels.

Other possible explanations may need to be explored to account for these findings. A first possibility is that the range of stimuli available in this experiment did not include a voice that could be clearly identified as having a relaxed quality. Some alternative explanation might also need considering, e.g., the possibility that these states (*relaxed-stressed*) are not expressed in voice in Japanese.

Relaxed: summary

- Striking cross-language differences.
- For the European languages, the trends are similar:
- *Relaxed* is signalled by lax-creaky, whispery and breathy voice.
- As expected, high dynamic f_0 does not contribute in signalling *relaxed.*
- For the Japanese, none of the stimuli signalled *relaxed.*

4.5.3 Stressed

Stress is related to increased sympathetic arousal. Acoustic manifestations of stress include higher concentration of high frequency energy (suggestive of tenser voice quality), high amount of voice perturbation (jitter, shimmer), increased f_0 and f_0 range, raised f_0 floor and maximum, greater intensity (higher vocal effort) and faster articulation (Douglas-Cowie & Cowie)[24]. Tense voice was found to be associated with *stressed* in Gobl and Ní Chasaide (2003b). The expectations here are therefore for the tense voice and the stimuli with high f_0 level and high f_0 dynamics to yield high ratings for *stressed.*

The 'VQ only' stimuli

The results show that tense voice yields the highest ratings, significantly so relative to modal, for all language groups tested except Japanese. Clear voice-to-affect mapping is found, which supports the descriptions in the literature.

None of the stimuli cued *stressed* for the Japanese listeners. As was found with the Japanese responses to relaxed, there was enormous interreater variability in this test (see Appendices 1 and 6).

The 'f_0 only' stimuli

Again, there is a striking cross-language difference in that the Japanese responses are quite unlike those of the three European languages.

Dynamic or high level f_0 is expected in the expression of stress. Note that modal + f_0 'indignation', modal + f_0 'fear' both have the highest f_0 contours and might therefore have been expected to be the most likely candidates for the expression of *stressed.* This was in fact

[24] *http://www.image.ntua.gr/oresteia/private/contributed_docs/reports/Voice-and-Stress.pdf.*

found: modal + f_0 'indignation' yields the highest ratings relative to modal for Hiberno-English, Russian and Spanish listeners. Modal + f_0 'fear' is considerably less effective, however (except for the Spanish subjects, values are subthreshold). We conclude that the additional dynamics in the modal+ f_0 'indignation' stimulus contributes importantly to its affective load.

For the Japanese subjects, none of the present stimuli were rated for the affect stressed, and again there was considerable interreater variability in responses (Appendix 6) suggesting a high degree of uncertainty with this test.

The combined 'VQ + f_0' stimuli

It is striking that either tense voice or the very high level and dynamically varied f_0 contour of f_0 'indignation' elicit high ratings for *stressed.* It would seem reasonable to expect that combining tense voice with f_0 'indignation' would yield even higher responses. This was not found generally to be the case. The addition of f_0 'indignation' to tense voice results in no increase in ratings, except for the Spanish listeners. For the Spanish listeners, it enhances the impression of *stressed* relative to tense, though not significantly.

Modal + f_0 'indignation' achieves the ratings that are as high as those of tense or tense + f_0 'indignation', in other words, here there is no additive effect.

Not surprisingly, none of the combined stimuli achieved high ratings for the Japanese subjects, and responses showed (again) low interrater agreement.

Cross-language similarities/differences

Stressed is another affect that revealed striking cross-language differences:

Tense voice or the high f_0 contour 'indignation' cue *stressed* to the three European languages tested but not to the Japanese. Hiberno-English, Russian, Spanish listeners largely agree in ratings, although there are minor differences in the relative strength of ratings. Spanish listeners may be slightly less sensitive to the tense voice cue than the Hiberno-English and Russian listeners.

The Japanese subjects rated none of the stimuli highly for *stressed.* Although the mean was around zero, there was considerable cross-subject variability and apparently uncertainty in these ratings.

The question that arises (as with relaxed) is whether these affects are cued by voice quality or f_0 cues. This is not a question that can be answered here. From the outset, one was

aware that the range of possibilities included here is not exhaustive and that particular voice qualities (and/or f_0 cues) might exist that are important for some of these affects, but are simply not included here.

As with *relaxed*, the post-hoc translation test (Section 3.3.2 of Chapter 3) explored the issue of the affective labels, and particularly whether the translations were appropriate. Our conclusion there was that it was unlikely that these cross-language differences reflect a lack of equivalence in the terms used.

It could be hypothesised that the expression of stress is suppressed by the display rules operating in the Japanese society (a collectivist society). For example, one could speculate that there are no stereotypical 'iconic' vocal expressions for either stressed or relaxed, if such expressions were to be socially unacceptable. This does seem, however, somewhat unlikely.

Stressed: summary

- Marked cross-language differences: Japanese vs. Russian, Spanish and Hiberno-English.
- None of the stimuli used in this study generated high ratings for *stressed* for the Japanese subjects.
- For the European languages, *stressed* is cued by tense voice, by f_0 'indignation' or by a combination of tense voice and f_0 'indignation'.
- Tense voice on its own is slightly less effective for Spanish than for Russian and Hiberno-English.

4.6 Sad-Happy

4.6.1 The statistical analyses

The results of statistical analysis were as follows. For the first part of the 'group' ANOVA, the three-way interaction (Target emotion group x Stimulus type x Language) was statistically significant, $F(9.2, 238.3) = 1.93$, $p = 0.047$, partial $\eta^2 = 0.07$. Two-way interactions Target emotion group x Language and Stimulus type x Language were not significant, but Target emotion group x Stimulus was $F(3, 238.3) = 76.13$, $p < 0.001$, partial $\eta^2 = 0.49$. The test of between-subjects effect was not significant: $F(3, 78) = 0.24$, $p = 0.87$, partial $\eta^2 = 0.01$.

When the design included the group 'indignation', the three-way interaction (Emotion x Stimulus x Language) was not statistically significant: $F(11.4, 296.1) = 1.8$, $p = 0.051$, partial $\eta^2 = 0.07$. While the two way interactions Target emotion group x Language and Stimulus type x Language were not significant, the effect of Target emotion group x Stimulus type interaction was found to be highly significant statistically $F(3.8, 296.1) = 40.9$, $p < 0.001$, partial $\eta^2 = 0.34$. The test of between-subjects effect was not statistically significant: $F(3, 78) = 0.32$, $p = 0.81$, partial $\eta^2 = 0.01$.

The results suggest that the difference in ratings was mainly due to the type of stimulus manipulation and the way the stimuli were grouped, but that the trends were similar across the languages tested.

Analysis of simple contrasts in which modal voice was compared to all other stimuli showed significant stimulus effect, $F(2.4, 190.4) = 76{,}9$, $p < 0.001$, partial $\eta^2 = 0.5$. Stimulus x Language interaction was not significant $F(7.3, 190.4) = 1.46$, $p = 0.18$, partial $\eta^2 = 0.05$. Between-subjects effect of language was not significant $F(3, 78) = 0.59$, $p = 0.62$, partial $\eta^2 = 0.02$.

The raters' agreement was as follows:

Hiberno-English:	ICC (r) = 0.44, ICC (R) = 0.94.
Russian:	ICC (r) = 0.61, ICC (R) = 0.97.
Spanish:	ICC (r) = 0.41, ICC (R) = 0.93.
Japanese:	ICC (r) = 0.43, ICC (R) = 0.94.
All:	ICC (r) = 0.48, ICC (R) = 0.99.

The raters' agreement was found substantial for the Russian listeners and moderate for the listeners from the other groups. The overall agreement was moderate. It can be concluded that the raters were coherent in assigning the ratings to the stimuli, and that no prominent cross-language differences are in evidence.

4.6.2 Sad

Sadness as a low activation state has been associated in the literature with breathy, whispery or creaky voice, flat f_0, monotonous pitch (low mean f_0, narrow f_0 range, little f_0 variability), low mean intensity, low speech rate and low precision of articulation (see summary Table 2.4 in Chapter 2). In Gobl and Ní Chasaide (2003b), sadness was found to be associated with lax-creaky voice.

The 'VQ only' stimuli

Lax-creaky voice was associated with sadness for all languages, as was whispery voice, but with near-threshold ratings. Relative to modal, the difference is significant for all languages suggesting that the voice quality is crucial for signalling this affect.

There is clear mapping of voice quality to affect, and contrary to the literature claims, lax-creaky rather than breathy proves the most effective. However, the difference in ratings between the two stimuli is not statistically significant. Note also that lax-creaky voice includes a considerable breathiness component.

Languages show a rather remarkable agreement here; the trends emerging are very similar.

The 'f_0 only' stimuli

The f_0 contour 'sadness' on its own is not effective in cueing *sad*, although there is a very weak shift towards *sad* in ratings relative to modal. Insofar as f_0 'boredom' is a low flat contour, one might have expected this to cue sadness, but this was not found to be the case.

The combined 'VQ + f_0' stimuli

The addition of f_0 'sadness' to breathy voice appears to slightly enhance the ratings for *sad* for all languages, although the difference does not reach statistical significance. Note that there is a slight enhancement also of whispery voice when combined with f_0 'fear'.

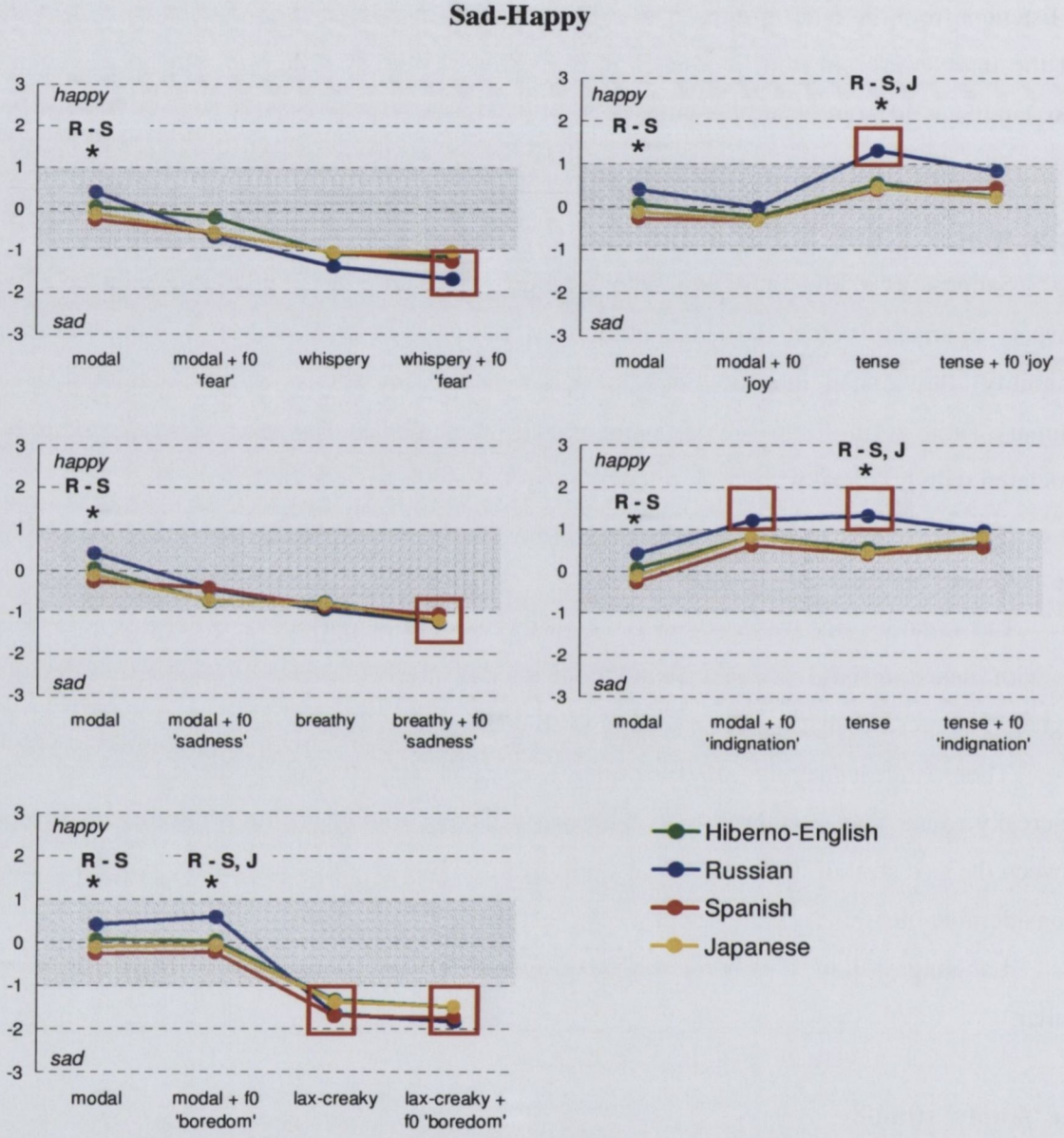


Figure 4.5. *Affective rating for each of the stimuli within the 'target emotion groups'. Asterisks show significant differences among language groups at $p < 0.05$.*

The most efficient stimuli for cueing sadness appear to be lax-creaky and lax-creaky + f_0 'boredom', the difference between them not being significant. Apparently, voice quality cues are sufficient and effective for affect signalling here.

Cross-language similarities/differences

There are very similar patterns of responses across language groups for all types of stimuli.

Sad: summary

- *Sad* is best cued by lax-creaky voice for all languages (and not by breathy as suggested in the literature), as well as by whispery voice.
- f_0 'sadness' on its own, that is, when voice quality is modal, is ineffective in cueing *sad.*
- No significant enhancement is achieved in affect signalling by combining f_0 'sadness' and breathy voice quality.
- There is a high degree of agreement across languages.

4.6.3 Happy

The best documented settings for joy are loud modal or tense voice quality, dynamic pitch variation, and supralaryngeal settings for smiling, which results in raising the frequencies of all formants (not modelled in the stimuli used here). Moreover, joy is associated with high mean f_0 and f_0 variability, wide f_0 range, medium to high mean intensity, high intensity variability, increase in high frequency energy, faster speech rate and high precision of articulation. Furthermore, it is well known that joy/happiness is notoriously difficult to pin down in terms of production or perception correlates. In face to face communication, the facial cueing (the smiling gesture) is clearly of paramount importance. In perceptual experiments, joy has been reported to be frequently confused with other high activation states such as anger (Elfenbein & Ambady, 2002b; Fónagy, 2001).

The 'VQ only' stimuli

Following the elusiveness of the acoustic correlates of this affect, expectations for the recognition of this affect are usually low. In this test, not surprisingly, *happy* was found to be

poorly signalled by the 'VQ only' stimuli available in the set. Tense voice yielded ratings above the threshold for Russian (but not for the other languages tested).

The 'f_0 only' stimuli

Dynamic f_0 variation is expected to enhance the impression of joy/happiness (but not in the expression of quiet content). Nonetheless, f_0 'joy' failed to cue *happy* in any of these languages. Note, however, that f_0 'indignation' (the most dynamic of the contours here) does trigger a shift towards *happy* in all four languages, even though it is only above threshold for Russian.

The combined 'VQ + f_0' stimuli

No enhancement of the ratings achieved by tense voice on its own can be observed in the ratings yielded by the combination of tense and f_0 'joy' or tense and f_0 'indignation'. It was hypothesised at the outset of the study that the addition of large f_0 perturbations to particular voice qualities would enhance the affective ratings of strong emotions. This was not the case here.

Cross-language similarities/differences

Quite similar patterns of responses can be observed across languages overall, in that they all yielded low ratings for *happy*, regardless of stimulus type. One exception to this is the case of Russian where above threshold ratings were obtained for the tense voice stimulus and for the 'f_0 only' stimulus modal + f_0 'indignation'.

Happy: summary

- *Happy* is poorly signalled by the stimuli in the set, but it is cued by tense, voice and modal + f_0 'indignation' for the Russian listeners.
- Poor performance of f_0 'joy'.
- In general no enhancement of the ratings was achieved by the addition of high level f_0 contours to tense voice.
- Similar cross-language patterns, with a slight bias for Russian in terms of the relative strength of affective ratings.

4.7 Scared-Fearless

4.7.1 The statistical analyses

The results of statistical analysis showed that the following two way interactions were found significant for the first part of the 'group' ANOVA: Target emotion group x Language, $F(4.9, 126.5) = 2.76$, $p = 0.02$, partial $\eta^2 = 0.1$; Target emotion x Stimulus type, $F(3, 238) = 63.8$, $p < 0.001$, partial $\eta^2 = 0.45$. Stimulus type x Language interaction was not significant $F(3.7, 97.4) = 1.19$, $p = 0.31$, partial $\eta^2 = 0.04$. The three-way interaction Target emotion group x Stimulus type x Language was not significant either: $F(9.2, 238) = 1.82$, $p = 0.065$, partial $\eta^2 = 0.07$. The test of between-subjects effect (Language) was not statistically significant: $F(3, 78) = 2.6$, $p = 0.06$, partial $\eta^2 = 0.09$.

When the design included the group 'indignation' the three-way interaction was statistically significant: $F(10.2, 265) = 2.81$, $p = 0.02$, partial $\eta^2 = 0.1$. The test of between-subjects effect (Language) was not statistically significant: $F(3, 78) = 1.72$, $p = 0.17$, partial $\eta^2 = 0.06$.

In the analysis of simple contrasts in which modal was compared to all other stimuli, significant Stimulus x Language interaction was found: $F(6.9, 178.4) = 2.4$, $p = 0.025$, partial $\eta^2 = 0.08$. Between-subjects effect of language was significant $F(3, 78) = 4.4$, $p = 0.007$, partial $\eta^2 = 0.14$.

The raters' agreement was as follows:

Hiberno-English:	ICC (r) = 0.46, ICC (R) = 0.95.
Russian:	ICC (r) = 0.55, ICC (R) = 0.96.
Spanish:	ICC (r) = 0.35, ICC (R) = 0.92.
Japanese:	ICC (r) = 0.31, ICC (R) = 0.90.
All:	ICC (r) = 0.39, ICC (R) = 0.98.

The raters' agreement was moderate for Hiberno-English and Russian groups and poor for the Spanish and Japanese, the overall agreement was poor/borderline moderate. The results suggest that some discrepancy existed in the way the stimuli were rated by the listeners; however, there are no prominent cross-language differences.

4.7.2 Scared

Fear is a high activation state, it has been associated with whispery and whispery/breathy falsetto voice (Burkhardt, 2009) as well as with high f_0 range, fast speech

rate, high mean f_0 and f_0 variability, wide f_0 range, jitter and fast speech rate (see Table 2.4 in Chapter 2). The addition of a high level f_0 contour is expected, therefore, to enhance the affective ratings.

The 'VQ only' stimuli

None of the 'VQ only' stimuli in the set has signalled *scared* to the listeners. There is a clear tendency for whispery to shift the ratings towards the *scared* end of the scale for all languages (but Spanish) relative to the modal voice stimulus (which has been rated by all listeners on the *fearless* end of the scale), but the magnitude of the shift is not sufficient to suggest effective cueing of *scared.*

No clear voice to affect mapping was found here. Descriptions in the literature suggest falsetto voice as signalling this affect. This rather extreme voice quality was not included in the set, and clearly it would be useful to include it in future work.

The 'f_0 only' stimuli

A clear contribution of f_0 is evident. Modal + f_0 'fear' is rated by all the languages as *scared* and the shift is substantial relative to modal (which gets ratings on the *fearless* end of the scale). Note, however, that the actual ratings for *scared* are near-threshold. This shift is evident not only for f_0 'fear', but for f_0 'sadness' and 'joy' However, ratings for the latter two contours are close to zero, and the shift relative to the neutral contour of the modal voice reflects more that modal voice has a more *fearless* quality.

In the case of the Spanish listeners, the f_0 'indignation' contour yields some degree of *scared*, something not found for the other languages.

The combined 'VQ + f_0' stimuli

While whispery on its own does not generate high ratings for *scared,* the addition of the f_0 contour 'fear' enhances the ratings significantly for all the languages. The additive effect of voice quality and f_0 contour is clear here, as can be seen when we compare the modal + f_0 'fear' stimulus and the whispery + 'f_0 fear' stimulus.

This result does support the initial hypothesis that the addition of major f_0 perturbations to particular voice qualities would enhance their affective colouring, for strong emotion.

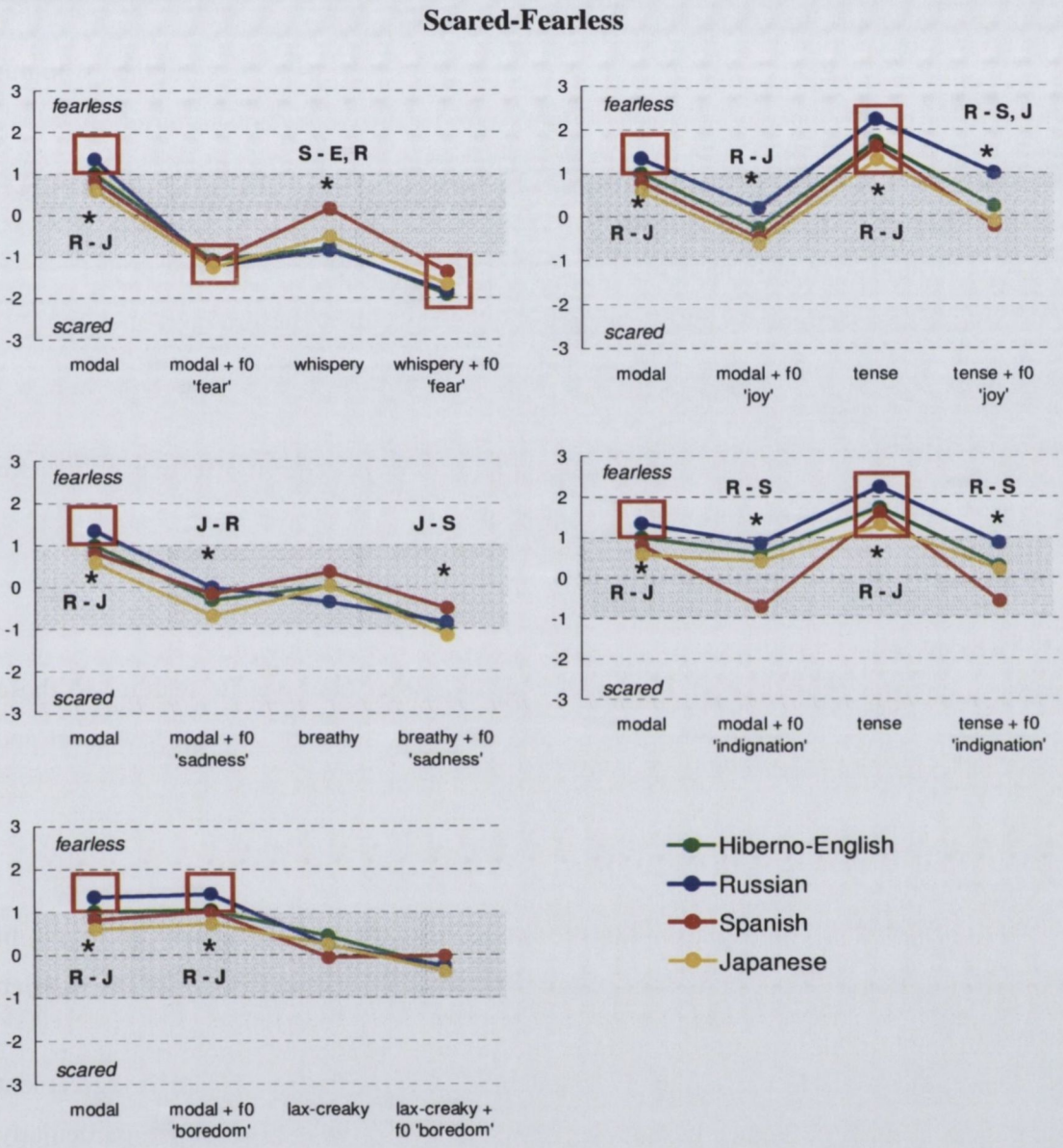


Figure 4.6. *Affective rating for each of the stimuli within the 'target emotion groups'. Asterisks show significant differences among language groups at $p < 0.05$.*

Cross-language similarities/differences

Very similar patterns emerge across the languages: in all cases the combined stimulus involving whispery voice and the f_0 'fear' contour proved effective in cueing *scared.*

Scared: summary

- Similar patterns emerge across the languages tested.
- The combined stimulus whispery + f_0 'fear' yields highest ratings.
- Strong emotion is better signalled by the combined stimulus.
- 'VQ only' stimuli not effective.
- The ' f_0 only' stimulus modal + f0 'fear' does achieve affective ratings for scared, but near threshold.

4.7.3 Fearless

For the vocal expression of *fearless* we would expect the opposite features to those associated with *scared*: tense to modal voice and an f_0 contour with a rather low level and dynamic range.

The 'VQ only' stimuli

Tense voice yields the highest ratings for *fearless* for all the language groups tested. In addition, *fearless* can be cued by modal (though the ratings are relatively high for the Russian listeners only).

Clear voice to affect mapping is found here, and the languages are in agreement although there is some difference in the relative magnitude of ratings (which are particularly high for the Russian listeners).

The 'f_0 only' stimuli

As expected, the f_0 contour which best conveyed the affect *fearless* was the one with the lowest and most flat f_0 contour, modal + f_0 'boredom'. Modal voice with the neutral f_0 yielded almost the same results. Note that only for Russian are values well above threshold. So in comparison to the 'VQ only' stimulus (tense voice) there is less clear affective signalling here.

The conclusion that the f_0 contour has to be rather flat for *fearless* to be communicated is also supported by the fact that stimuli with dynamic, high level f_0 contours (e.g., 'joy', 'indignation') yielded lower ratings for *fearless*. In the case of f_0 'fear' this leads to a clear signalling of the opposite affect *scared.*

The combined 'VQ + f_0' stimuli

The combined stimuli do not lead here to any enhancement of the *fearless* ratings obtained with tense voice. In view of the responses just discussed for the 'f_0 only' stimuli this is not at all surprising. In this experiment tense voice was combined with dynamic, high level f_0 contours 'joy' or 'indignation', and clearly these f_0 contours militate against the cueing of *fearless*.

The relative ineffectiveness of the combined stimuli in signalling *fearless* points clearly to the combination not being optimal for this affect.

Cross-language similarities/differences

Similar patterns of responses from different language groups are evident overall. All languages see tense voice as signalling *fearless*.

Fearless: summary

- *Fearless* is signalled by tense voice for all the language groups tested.
- Although not as potent as the 'VQ only' tense voice stimulus, the low flat f_0 contour 'boredom' achieved higher affective ratings than the high dynamically varying contours.
- Overall, a good deal of cross-language agreement can be observed.

4.8 Summary and conclusions

The aim of this study was to investigate how particular voice qualities map to affect in four different languages and to what extent f_0 contours with different level and dynamics can enhance this association. The results point, although for a limited selection of languages, at what might be universal in voice cues to affect mapping. The results further draw attention to possible cross-language/cultural differences in the way the voice is used to communicate affect.

I will sum up the findings in terms of the research questions posed at the introduction to this chapter. Figures 4.7 – 4.8 are included to help us get a quick overview of where there is cross-language agreement in the mapping of affect to these individual stimuli. In these figures, the mean ratings obtained for all stimuli are shown for each subtest. Red boxes draw attention to the cases where all four languages concurred in rating a particular stimulus as having a clear affective colouring, taking the +/-1 value as the threshold, as in earlier figures.

Similar figures are included in Appendix 7, which expand on Figures 4.7 and 4.8 by including for each of the subtests representations that allow us to look at not only cross-language agreement (in red boxes as above) but also: cases where three languages show above threshold responses for particular stimuli (blue boxes); cases where two languages show above threshold responses for particular stimuli (grey boxes); or even a single language (orange boxes). These latter figures in Appendix 7 are useful in allowing us to focus on the prevalence of cross-language differences.

Q1: How do 'VQ only' stimuli map to affect?

All non-modal 'VQ only' stimuli get associated with a number of affects, although languages may vary in the choice of stimulus to affect mapping and in the strength of the affective rating accorded to a particular stimulus. See summary data in Figures 4.7 – 4.8 as well as the more elaborated data in Appendix 7.

Lax-creaky and tense voice proved to be effective in the following general ways: lax-creaky voice signals *bored* and *sad,* tense voice signals *fearless* and *indignant* for all the languages tested.

In most cases, there is no one-to-one voice-to-affect mapping, and the same voice can be associated with a number of affects. For example, whispery voice could signal not only *intimate* (as suggested in the literature), but also *apologetic, bored* and *relaxed* (for Hiberno-English and Russian). This is in keeping with the findings in earlier studies (Campbell, 2004c; Gobl & Ní Chasaide, 2003a), where it is clear that there is no one-to-one mapping between voice quality and affect.

Similarly, the same affect can be cued by more than one voice quality. For example, in this data set, *relaxed* has been cued by whispery, breathy and lax-creaky voice (for the European languages tested).

Voice quality on its own (that is, without f_0 variation) is a poor indicator of the affects *scared* and *interested.* (This holds at least for the set of voice qualities available in this test.)

Modal voice, used in this study as a neutral baseline, was not expected to be associated with any affect. However, it does appear to be associated with *fearless* and *formal* by the speakers of one language - Russian. This may be a pointer to possible different neutral (habitual baseline) voice quality in these languages.

Q 2: Does the addition of a dynamic f_0 contour enhance affective rating of a particular voice quality stimulus? The hypothesis here is directional: the research is predicting that f_0-enhanced stimulus from the 'VQ + f_0' series will be more effective in signalling affect, particularly strong emotions, than the corresponding stimulus from the 'VQ only' series.

If we look at the summary information in Figures 4.7 and 4.8 and the more detailed elaboration for all subtests in Appendix 7, it is clear that overall, the combined stimuli were very frequently generating above threshold affective responses.

Two combined stimuli were particularly effective: f_0 contour 'fear' with whispery voice and f_0 'indignation' with tense voice. These substantially enhanced the ratings for such affects as *scared, apologetic, interested* for all the language groups tested (See Figures 4.7 and 4.8).

One combination of voice quality and f_0 – tense + f_0 'indignation' failed to signal the intended affect. For all languages tested, the addition of f_0 'indignation' did not (against expectations) enhance the impression of *indignant* relative to tense but caused the opposite effect, lowering the rating in particular for Spanish and Japanese listeners.

Q3: To what extent can the f_0 contour (alone, i.e. with modal voice) induce the same affective rating?

Different f_0 contours available in the range showed different potency in affect signalling. It is evident that modal + f_0 'fear' and modal + f_0 'indignation' were consistently associated with at least two affects (*scared* and *interested* respectively) common to the listeners from all the four language groups.

Note that modal + f_0 'fear' although effective in signalling *scared*, was not at all as effective as the combined stimulus whispery + f_0 'fear'. In the case of modal + f_0 'indignation', it was slightly less effective than the combined stimulus tense + f_0 'indignation', for some languages, but the difference was not statistically significant.

The f_0 contours modal + f_0 'sadness' and modal + f_0 'joy' 'failed' to communicate any affect to the speakers of any language tested. Of all the contours available, the listeners therefore favoured the contours with the highest f_0 level and dynamics, as marking affect.

The lack of contribution from f_0 'boredom' was somewhat expected as it had minor differences relative to the neutral f_0.

Q4: What stimuli (other than 'VQ only' stimuli) are consistently associated with a particular affect?

Overall, for the range of stimuli included in this experiment, those which incorporate distinct voice qualities (whether 'VQ only' or 'VQ + f_0') emerged as being the most effective in signalling affect. The 'f_0 only' stimuli are relatively less effective, except the affects *interested, scared* and *stressed.* This holds for all the languages tested. This can be ascertained by looking at the summary of results in Figures 4.7 - 4.8 and at the elaborated summaries in Appendix 7.

Q5: To what extent do we find agreement among the languages, to what extent do distinct language-specific trends emerge?

Languages agree

As can be seen from the broadly similar trends in results for the different languages in Figures 4.1, 4.2, 4.5 and 4.6, there is considerable cross-language agreement in the way these stimuli signal quite a number of affects. There are of course some differences in the strength of affective ratings. This cross-language agreement is seen for the affects representing

emotions proper (*sad, scared, indignant*), and also attitudes and interpersonal stances (*apologetic, interested, bored, fearless*).

The following stimuli were similarly rated by listeners from all these language groups, (a kind of 'universality'), with differences showing up in the strength of the affective rating. These cases are highlighted in red in Figures 4.7 and 4.8:

a. whispery + f_0 'fear' signalling *apologetic* and *scared*,
b. lax-creaky and lax-creaky + f_0 'boredom' signalling *bored* and *sad*;
c. tense signalling *indignant* and *fearless*;
d. tense + f_0 'indignation' and modal + f_0 'indignation' signalling *interested;*
e. modal + f_0 'fear' signalling *scared.*

In other words, there is agreement in the selection of the stimulus to represent the same affect, yet with a clear difference in the strength of affective ratings demonstrating difference in the relative potency of the stimulus in cueing the affect for a particular language.

Languages disagree

In two subtests, particularly, the cross-language difference was substantial. These subtests involved the *intimate-formal* and the *relaxed-stressed* pairs of affects.

The *intimate* affect was associated for Spanish and Japanese subjects with tense/modal voice combined with f_0 'indignation', a stimulus which was associated rather with *formal* for the Hiberno-English and Russian subjects. On the other hand, *intimate* was cued by lax-creaky voice and lax-creaky + f_0 'boredom', whispery voice and whispery + f_0 'fear' for Hiberno-English and Russian, - qualities that did not evoke intimacy for the Spanish and Japanese listeners.

There was a conspicuous gap in the cueing of *formal* with any of the stimuli presented here for the Japanese listeners.

There was no affective response from the Japanese in the *stressed-relaxed* subtest for any of the stimuli presented in this experiment.

The differences noted in voice to affect association, are likely to be linked to underlying cross-cultural differences. Note that the major differences that have shown up here involve mainly attitudes or interpersonal stances (rather than emotions as such): *intimate-formal, relaxed-stressed*, which are more likely to reflect culture-shaped learned behaviour. This is an area we can only speculate on. There may be culture-specific differences in the display rules for certain affects which would affect the extent to which there would be iconic

voice-to-affect associations. For example, it could be that none of the stimuli available signalled *stressed* or *relaxed* for the Japanese because the expressions of stress are frowned upon by this collectivist culture and are therefore not expressed through vocal cues. The same is true for *formal*. The expression of formality is done in Japanese through the system of honorifics *keigo*, and vocal expression may very well be only of minor relevance for this affect.

One issue that must also be borne in mind is the possibility that the affective labels are not equivalent for the different languages. This was an issue that was considered and discussed with the translators when the test was being designed, and we were aware of some (perhaps slight) unavoidable non-equivalence of the terms for *intimate* in Spanish as compared to English. Following the implementation of the cross-language study, the issue of equivalence of labels was revisited for the Japanese terms, particularly as the author has no familiarity with this language (being fluent in the other three). A post-hoc test was conducted for Japanese to see whether the labels used were optimal and to get some sense of where labels might not have full semantic equivalence.

The results of this test suggest that the *relaxed-stressed* or indeed the *formal* term were not in themselves problematic. However, for *intimate* some non-equivalence did emerge, which mirrors the situation for Spanish. While the English and the Russian translation for the term 'intimate' can be associated not only with decreased social/physical distance, closeness, familiarity, and possibly with sexual overtones, the Japanese labels (like the Spanish) does not include the possible sexual overtones. Intuitively, we did not feel that the semantic differences were likely to explain the very different voice to affect mappings emerging in the present experiment. However, this issue does suggest some interesting future research where one could try to elicit how stimuli such as these map to a richer coverage of affective states.

Another factor that could contribute to cross-language differences such as those observed here would be the neutral voice (baseline) prevalent in these different languages. To my ear, the neutral voice of Spanish (and indeed Japanese) is very different from that of English or Russian. If the neutral voice in language A is different from language B and if the neutral voice is similar to a quality which has affective overtones in language B, then cross-language differences have to be expected in the voice to affect association for these languages. We cannot comment on the neutral voice of these languages, but this is an interesting area in itself which of course intersects with how voice is exploited in a particular language to signal affect.

Discussion so far has concerned the very dramatic differences found in the *relaxed-stressed* and *intimate-formal* subtests. Further more subtle differences also surfaced in some of the other subtests. For example, languages may differ in the relative sensitivity to voice quality or f_0 cues. In particular, Japanese subjects appear to be more sensitive to f_0 cues for the cuing of *apologetic* and *interested*, compared to the other languages where the voice quality cues appear to be key.

One further, relatively small cross-language difference concerned language-specific preferences, even when there is agreement in the broad quality to affect mappings. For example, *apologetic* is signalled by lax-creaky and whispery voice Russian, Hiberno-English and Spanish. However, the lax-creaky quality is the more effective in Russian and Spanish, whereas for Hiberno-English whispery voice is considerably more effective.

As pointed out in the earlier chapters, the stimuli presented in this experiment are necessarily a very limited subset of the infinite variety that exists in real life. In the case of voice qualities, not only have certain qualities been omitted that are often mentioned in the literature (harsh voice and falsetto), but extreme versions of these qualities have not been included. In the case of the f_0 contours, although it was felt that we were including a reasonable sampling of the types of differences in f_0 level and f_0 dynamics that are often mentioned in the literature, it is clear that these are also a limited sampling of the endless possibilities that exist, including the possibility of varying the shape of the f_0 contour (e.g., so that the final accent is a rising rather than a falling one). Finally, the same points have to be made about the combined stimuli.

Where, as with intimacy, the affects are strongly signalled, but the different languages appear to use very different qualities in the signalling, we can be fairly certain that we are dealing with a major difference in voice to affect mapping, regardless of the factors that might contribute to this difference.

Where, as with the Japanese responses to *stressed/relaxed* and to *formal*, there is a gap in the affective signalling so that no affective response appears in the experiment, the conclusions have to be rather different. In these cases it is possible and indeed likely that the gaps arise because the stimulus set did not include the qualities that would be used in this language for the signalling of these affects. Here, one would need eventually to look at production data such as might emerge from spontaneous speech corpora. From such data might emerge some further voice qualities (and f_0 contours) that would be more appropriate and which could be tested within the kind of perception experiment presented here.

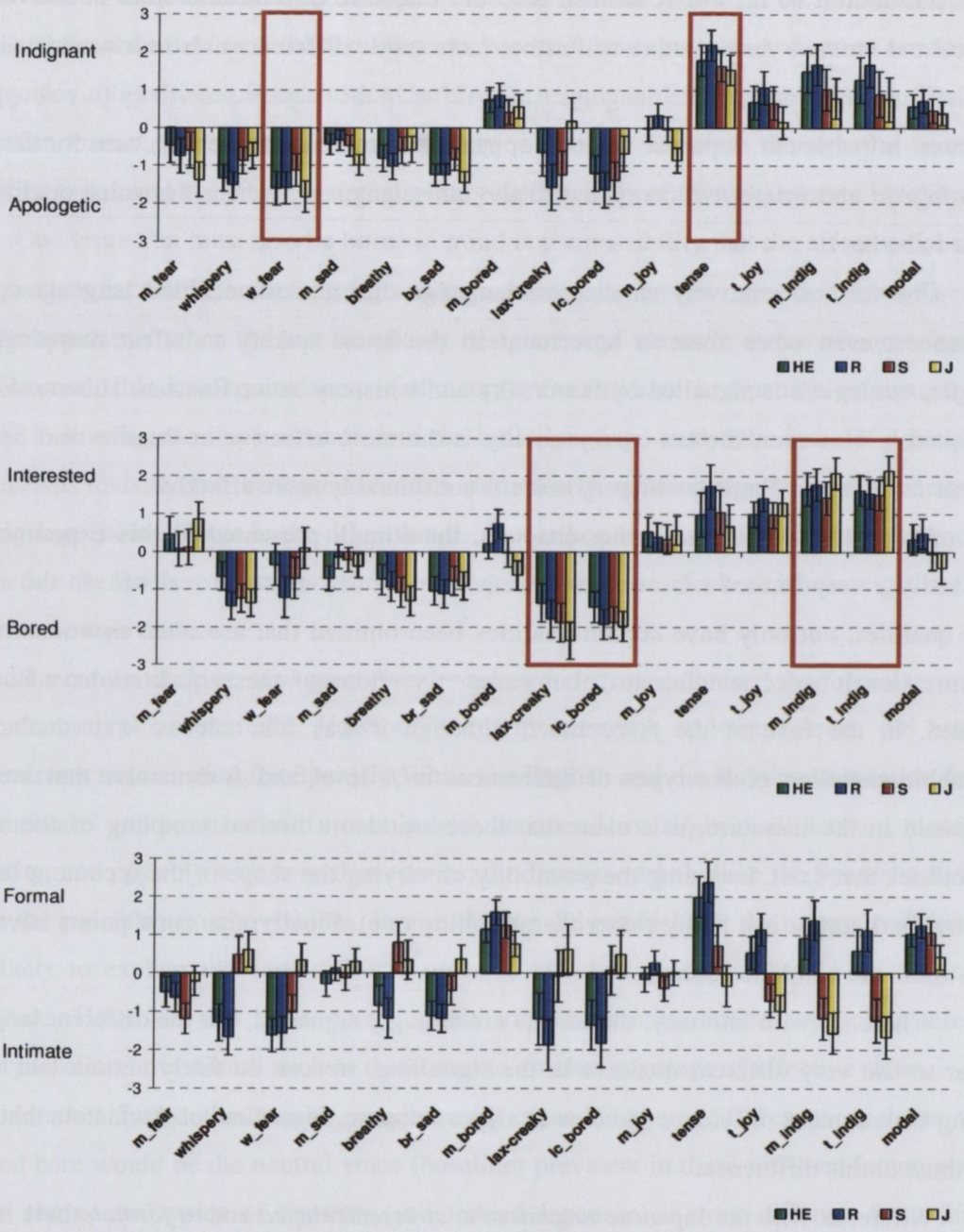


Figure 4.7. *Mean ratings obtained by different stimuli: cross-language comparison. Voice-to-affect associations similar across language groups are highlighted.*

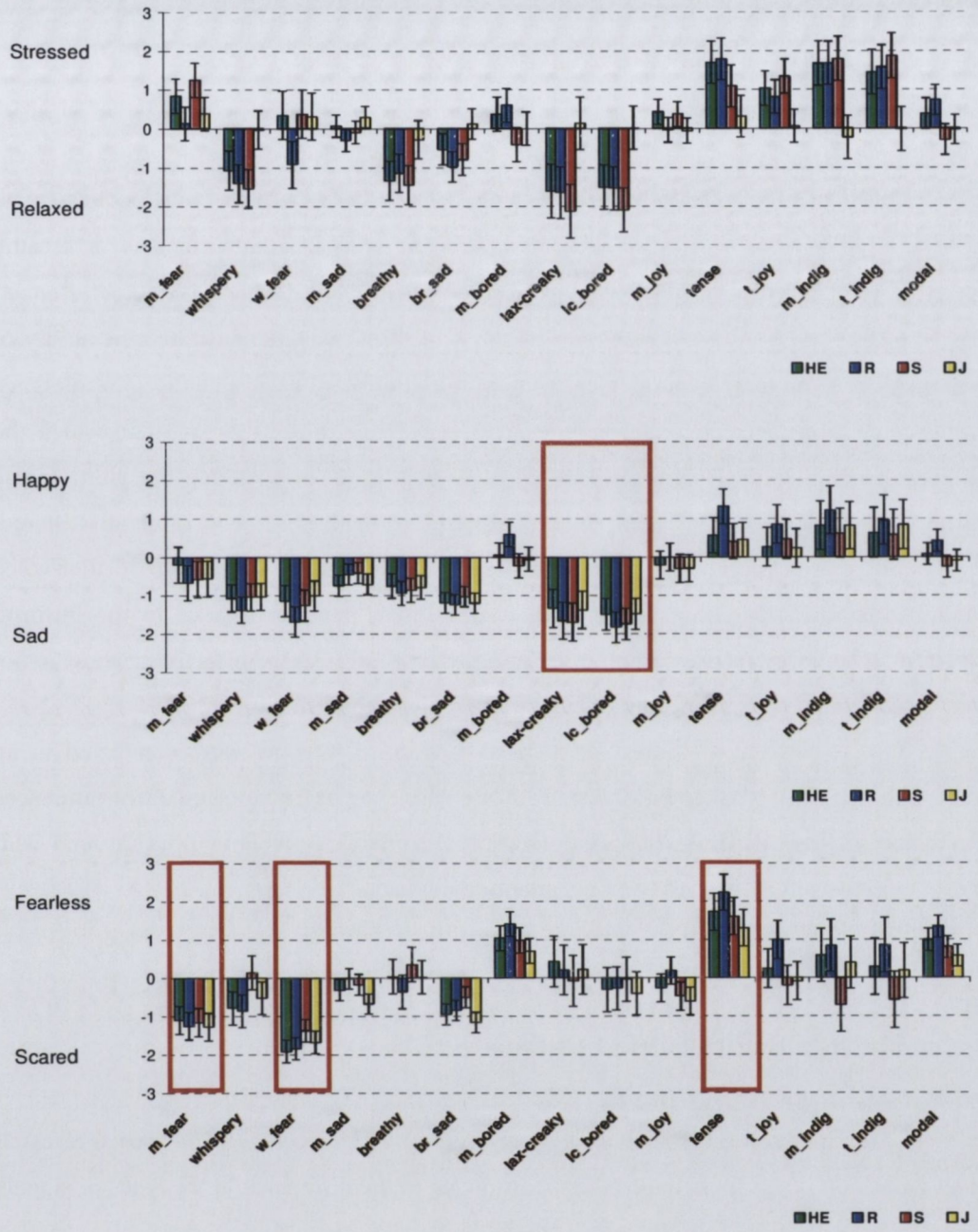


Figure 4.8. *Mean ratings obtained by different stimuli: cross-language comparison (continued). Voice-to-affect associations similar across language groups are highlighted.*

5 Chapter five
Loudness and voice quality

5.1 Introduction

Different voice qualities are characterised by different perceived loudness, e.g., tense and harsh voice will most likely be perceived as louder than whispery or breathy voice. Experiments reported in Gobl and Ní Chasaide (2003b), Gobl (2003), Gobl, Bennett and Ní Chasaide (2002), as well as in Chapter 4 of this thesis demonstrated the ability of voice quality variation to communicate emotion and affect. In these studies, an utterance synthesised with different voice qualities was tested to explore the extent to which the different voice qualities might alter its affective colouring. In changing the glottal pulse shape to synthesise the different voice qualities, the loudness is concomitantly altered. Thus, while these experiments reported distinct affective associations with particular voice qualities, the question arose as to whether this might partially be due to the intrinsic loudness differences in the stimuli. Our initial hypothesis is that affective cueing is not simply a consequence of the loudness variation in these voice quality stimuli.

Two experiments outlined in sections 5.3 and 5.4 below were conducted in an attempt to clarify to what extent the intrinsic loudness of certain voice qualities influences perception of their affective colouring. Section 5.2 provides theoretical background with regard to perceived loudness as a voice quality component.

5.2 Loudness: theoretical background

5.2.1 Definition of loudness

Loudness has been studied extensively both in psychoacoustics and speech research. Their definition of loudness may differ, mainly due to the difference in the types of signals studied.

Loudness is generally defined as the perceptual property of a sound: the perceived strength of a sound. In other words, perceived loudness is a global attribute of sound that is related to the magnitude of the auditory sensation (Plack & Carlyon, 1995). Scharf (1978, p.187) defines loudness as the 'subjective intensity of a sound'. According to this definition, in the perception of loudness the listener must respond to the magnitude of sound in a conscious way. Consistent responses obtained from a sufficiently large group of

listeners in the course of controlled experiments form the basis for describing the relationship between perceived loudness and variable physical properties of sound.

In his definition of loudness, Moore (2003) uses the notion of loudness scaling and describes loudness as 'that attribute of the auditory sensation in terms of which sounds can be ordered on a scale extending from quiet to loud' (p. 127).

Perceived loudness corresponds most closely to the physical measure of sound intensity. It furthermore depends on such factors as the properties of the signal (spectral content and bandwidth or duration and intermittency of sound) and the conditions in which the sound is presented to the listener (for example, the background). There also exists important interaction between the properties of the signal and the listener. As pointed out by Scharf (1978), 'loudness resides in the listener, not in the stimulus' (p. 188). Perceived loudness will depend to various degree on factors such as stimulus presentation (binaural or monaural), whether the listener has been exposed to noise; whether the listener has a hearing impairment; and to what extent listening is a conscious process (Moore, 2003; Plack & Carlyon, 1995; Scharf, 1978; Verhey & Kollmeier, 2002; Zwicker & Fastl, 1999). The study of the perception of loudness and the way it is related to the temporal and spectral properties of a signal is fundamental to the understanding of the way in which the sounds are represented in the auditory system (Moore, 2003).

5.2.2 Measuring loudness

The fundamental complexity in measuring loudness lies in the fact that it is subjective auditory sensation, residing in the listener, and as a result cannot be measured directly. Assumptions about loudness perception have to be based on the results of controlled listening tests and are not always uncontroversial (Moore, 2003; Plack & Carlyon, 1995; Scharf, 1978; Zwicker & Fastl, 1999). Results obtained in the course of listening tests serve as a basis for development of objective instrumental techniques such as loudness models and loudness meters.

SUBJECTIVE MEASUREMENTS: LISTENING TESTS AND LOUDNESS SCALING

The main types of listening tests used in the study of loudness are loudness matching, magnitude estimation and magnitude production.

In loudness matching, the listeners are required to adjust the intensity of a test stimulus until it matches the loudness of the reference stimulus. In even simpler experimental methods, the listeners could be asked to choose the loudest stimulus out of the pair of stimuli presented or to rank the loudness of more than two stimuli presented

together. The method has proved particularly useful in testing the correlation between the frequency and bandwidth of a signal and its perceived loudness. In magnitude estimation, the listeners assign a number to the sound that is presented to them according to its perceived loudness or express, for a pair of stimuli, the difference between the perceived loudness of the test stimulus and the reference stimulus in terms of an interval or ordinal scale. In magnitude production, the listeners adjust the intensity of a test sound using the volume or gain control to match a specified loudness expressed as a number. They could also be asked to match the loudness of a test stimulus to a given reference sound so that the ratio between the perceived loudnesses of the two sounds corresponds to a certain number given by the experimenter, for example twice as loud (Moore, 2003; Plack & Carlyon, 1995; Skovenborg & Nielsen, 2004; Skovenborg, Quesnel, & Nielsen, 2004).

According to Plack and Carlyon (1995), magnitude production or magnitude estimation experiments assume that listeners can use numbers in a consistent and linear way. However, this is not always the case, as listeners' estimation of loudness may be biased by a variety of factors such as the range of stimuli; the first stimulus presented; the instructions given to the subject; the range of responses allowed; the symmetry of the response range (listeners tend to choose numerical responses in the centre of the intensity range available) and others. Due to large individual differences observed, consistent results can only be obtained by averaging judgments of a large number of subjects (Moore, 2003; Plack & Carlyon, 1995).

S. S. Stevens (1957) used the methods of magnitude estimation and magnitude production to develop loudness scales describing the relationship between intensity (the physical magnitude of sound) and perceived loudness. The perceived loudness can be expressed in sone.

Sone as a unit of perceived loudness was introduced by S. S. Stevens (1957). One sone is defined arbitrarily as the loudness of a 1 kHz pure tone with a level of 40 dB SPL presented binaurally from a frontal direction in free field. A sound judged by listeners in experiments based on magnitude estimation to be twice as loud as the 1 kHz 40 dB tone would have a loudness of 2 sones, a sound judged to be half as loud would have a loudness of 0.5 sones, and so on (Moore, 2003). The sone scale is a ratio scale, so the value expressed in sone is directly proportional to the perceived loudness (e.g., a sound that is perceived as twice as loud as a reference sound of 1 sone will have a loudness of 2 sones) (Skovenborg & Nielsen, 2004).

EQUAL LOUDNESS CONTOURS: LOUDNESS LEVEL

A more accurate way (relative to magnitude estimation) to describe the relationship between intensity and perceived loudness is loudness level. The loudness level measure was introduced by H. Barkhausen in 1920s to characterise the loudness sensation of any sound and is measured in phon. Loudness level of a sound in phon is defined as the level in dB SPL of a 1 kHz tone to which it sounds equal in loudness. That is, any tone that has the same loudness as a 40 dB SPL 1 kHz tone has, by definition, a loudness of 40 phon. When a sound has a loudness level of 50 phon, it means that it sounds as loud as a 50 dB SPL 1 kHz tone (Plack & Carlyon, 1995; Zwicker & Fastl, 1999). The phon scale is similar to the dB scale. When the loudness of two sounds is compared by adjusting the relative level to match their loudness, the difference is typically expressed in phon or dB (Skovenborg & Nielsen, 2004).

The relationship between loudness expressed in sones and loudness expressed in phon is given in Figure 5.1.

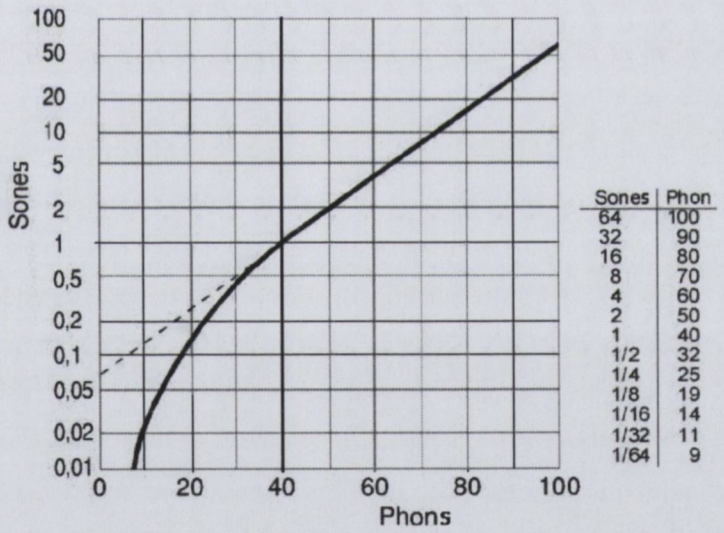


Figure 5.1. *The relationship between loudness in sones and loudness level in phon for a 1 kHz tone.*

Although loudness level can be measured for any sound, the best studied are the loudness levels for pure tones of different frequencies known as equal loudness contours (Figure 5.2).

Equal loudness contours can be obtained by asking the subjects to adjust the level of a 1 kHz tone until it has the same loudness as the test sound. An equal loudness contour is a line joining the levels of tones of different frequencies that have the same loudness in phon. By definition, all curves have to go through the sound pressure level at 1 kHz that

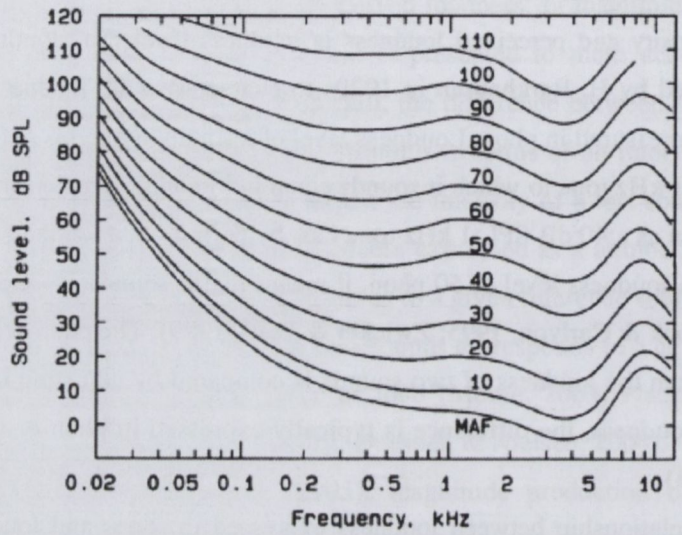


Figure 5.2. *Equal loudness contours representing the variation of loudness as a function of frequency (from Robinson & Dadson, 1956). MAF = Minimum Audible Field, absolute threshold curve.*

has the same value in dB as the parameter of the curve in phon: the equal loudness contour for 40 phon has to go through 40 dB at 1 kHz. Equal loudness contours hold for durations longer than 500 ms (Moore, 2003; Plack & Carlyon, 1995; Zwicker & Fastl, 1999).

Equal loudness contours are not parallel and change shape with increasing level: the large differences in SPL between equally loud low and medium frequencies decrease at higher level, therefore loudness does not change with sound pressure in the same way for all frequencies. With the increase of sound pressure level, loudness level grows at a faster rate for low frequencies than for middle frequencies. The shapes of equal loudness contours have important implications for the transmission of sound and have been used in sound level meters which are designed to give fairly accurate estimation of the perceived loudness of complex sounds (Moore, 2003; Scharf, 1978).

The shapes of the equal-loudness contours depend on the method used to measure them and vary markedly across studies. For example, free-field contours differ noticeably from the earphones listening at frequencies above 1000 Hz, and equal loudness contours for a plane sound field and diffuse sound field (in which the sound comes from different directions) are different (Moore, 2003; Scharf, 1978; Zwicker & Fastl, 1999).

Equal loudness contours for pure tones demonstrate an important dependence of perceived loudness on frequency; however, they cannot be regarded as indicators of subjectively perceived loudness. The sensation of loudness depends on other factors such

as bandwidth and frequency content of the signal as well as its duration (Zwicker & Fastl, 1999).

OBJECTIVE MEASUREMENT PROCEDURES: LOUDNESS MODELS AND LOUDNESS METERS

Objective loudness measurements include the application of loudness models that allow estimating the perceived loudness of the signal from its spectral characteristics. Loudness models fall into two categories: single-band and multi-band.

Single-band models are based on the assumption that loudness perception depends on the frequency of the signal. A frequency weighting (A or B weighting curve) is used to weight the intensity at each frequency to approximate to one of the equal loudness contours. Single-band models are used to estimate the loudness of broadband signals with similar spectral content, and within a relatively narrow level range (Skovenborg & Nielsen, 2004).

The multi-band models divide the signal into frequency bands that are subsequently combined into an estimate of perceived loudness. Multi-band models incorporating several frequency bands became possible with the discovery and calculation of spectral loudness summation and provide some fundamental advantages in the accurate modelling of perceived loudness (Skovenborg & Nielsen, 2004; Zwicker, Flottorp, & Stevens, 1957). Zwicker and Fastl (1999) argued that their model is based on the fundamental assumption that 'loudness is not produced from... the spectral distribution of the sound directly, but... is the sum of specific loudnesses that are produced at different critical-band rates' (p. 223). They further argue that 'transforming the physical spectra (level versus frequency) into specific loudness versus critical band rate yields the best hearing equivalent psychoacoustical values' (Zwicker & Fastl, 1999, p. 223). Zwicker's highly influential multi-band loudness model is described in ISO 532 standard as method B (ISO, 1975). The input to the model is the 1/3 octave spectrum of the sound, which is a good approximation to the critical bands and the output is the estimated loudness in sone (Plack & Carlyon, 1995).

Loudness models have been widely used in loudness meters and noise control devices. An overview of performance of both well established and relatively recent loudness models in calculating loudness of speech and music is given in (Skovenborg & Nielsen, 2004; Skovenborg, Quesnel, & Nielsen, 2004).

Loudness of complex sounds can be estimated using loudness meters. The intensity of the complex sounds is first weighted for each frequency according to the shape of the equal loudness contours, and then summed across frequencies. The shapes of the loudness contours are approximated by using 'A', 'B' or 'C' weighting network. For example, the 'A' weighting, which is roughly based on the 30-phon equal loudness contour, is used for

low sound levels, when low frequency components contribute little to the overall loudness of a complex sound. The 'C' weighting network is used for high level sounds, where the contribution of all frequencies to the overall loudness is similar. Although loudness meters provide a tool for comparing loudness of complex sounds, they do not give direct estimation of subjective sensation (perceived loudness). Their use is restricted to steady sounds of relatively long duration in which the separation of frequency bands is relatively narrow (Moore, 2003).

5.2.3 Factors influencing loudness perception: properties of the signal

Intensity

Loudness is the sensation that most closely corresponds to the sound intensity and is described as a monotonic function of sound intensity according to Stevens' power law. S. S. Stevens suggested that for pure tones the relationship between the loudness and intensity can be expressed as $S = kI^{0.3}$, where S is loudness, I is intensity, and k is a constant dependent on the listener and the units employed (Stevens, 1957). Expressed in terms of amplitude or sound pressure, the loudness of a given sound is proportional to its RMS pressure raised to the power of 0.6. Consequently, a relative change in loudness is proportional to a relative change in intensity so that, for example, a 10 dB increase in intensity corresponds to approximately a twofold increase in loudness. This power law relationship between loudness and intensity holds for sounds above about 40 dB SPL. For levels below 40 dB SPL, loudness changes more rapidly (Moore, 2003; Plack & Carlyon, 1995; Scharf, 1978; Zwicker & Fastl, 1999).

The loudness function is based on magnitude estimation. The standard loudness function is defined specifically for a 1 kHz pure tone and will therefore be different for tones of different frequency and for complex sounds. The shape of the loudness function will change depending on the duration of the tone, the background against which it is presented as well as on the properties of the auditory system of the listener. The loudness function at frequencies other than 1 kHz can be calculated from equal loudness contours (Scharf, 1978).

Scharf (1978) emphasises that it would be incorrect to identify loudness as subjective intensity (subjective in that it involves conscious appreciation of the stimulus) with physical intensity. Their close relationship, however, allows defining loudness as that attribute of a sound that changes most readily when the sound intensity is varied. He further summarises the difficulties inevitably encountered by the experimenter studying perceived loudness: 'Any sound, even one as simple as a pure tone, is more than just

louder or softer. People can respond to its pitch, size, density, duration, vocality, and annoyingness. A complex sound... has still more attributes: meaning, timbre, roughness, intermittency, and colour. Telling a subject to judge loudness and ignore all other attributes of a sound leaves the experimenter who attempts to measure an equal loudness contour or a loudness function at the mercy of the subject's interpretation of loudness (or intensity or strength). ...The experimenter has no way of determining that a judgement is right or wrong: only the listener can say what the ratio is between two sounds. The experimenter hopes that the subject chooses the right criteria. (The experimenter also hopes that the subject listens, not only to the instructions but to the stimuli. A subject may report a sound as much softer - even nonexistent - when he pays no heed to it than when he listens carefully. In the laboratory, getting subjects to pay attention to the stimuli is seldom a problem, but in the real world of intrusive noise, attention and habituation are often critical variables.)' (Scharf, 1978, pp. 228-229).

SPECTRAL EFFECTS

Subjective perception of the loudness of a signal depends on its spectral characteristics. According to Scharf (1978), the loudness of a pure tone depends, along with its intensity, on its frequency. The loudness of a complex sound depends on overall intensity, on the frequency of its components, and also on the frequency separation defined as the distance between the component with the lowest frequency and that with the highest frequency.

The perceived loudness of a sound depends on the frequency range it covers: the greater the frequency range covered by a sound the louder it is. For example, the loudness function for pure tones (the narrowest type of signals in terms of spectral distribution) and white noise (for which the spectral distribution is the broadest as the spectrum is continuous and flat, that is containing energy at all frequencies) will be different (Scharf, 1978; Zwicker & Fastl, 1999).

Skovenborg and Nielsen (2004) comment on spreading, a complex way in which signals of different frequencies interact. Because of spreading, the overall loudness of the signal is influenced by not only the actual frequencies of a signal, but also the higher and lower neighbouring frequencies.

BANDWIDTH, CRITICAL BAND AND SPECTRAL LOUDNESS SUMMATION

An important relationship exists between the loudness of a signal and its bandwidth. The loudness of a complex sound increases with its bandwidth even though its intensity is kept constant. The bandwidth shows practically no effect on loudness until a

certain critical bandwidth is reached, but as the bandwidth increases above the critical band, the loudness grows significantly with it. This phenomenon is referred to as (spectral) loudness summation: for sounds which excite more than one critical band the loudness is determined for each critical band separately and then summed across frequency. This sum yields larger loudness values than the loudness of a sound with the same intensity, but with a bandwidth lower than the critical (Moore, 2003; Skovenborg & Nielsen, 2004; Verhey & Kollmeier, 2002).

As evident from Figure 5.3, the loudness of a complex sound is practically independent of bandwidth at low sensation levels (10-20 dB SPL). At higher sensation levels up to between 40 and 60 dB, loudness increases more rapidly as a function of bandwidth. Loudness summation is greatest at moderate levels (Moore, 2003; Scharf, 1978; Zwicker & Fastl, 1999; Zwicker & Scharf, 1965).

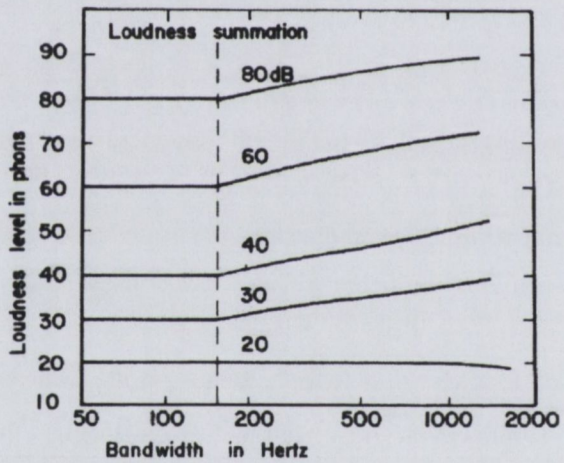


Figure 5.3. *Loudness level of a band of filtered white noise plotted as a function of its bandwidth. The centre frequency is 1 kHz (from Scharf, 1978).*

For multitone complex sounds, loudness level depends on the spacing of the components with respect to critical bands as well as on the loudness of individual components. Thus, the loudness level of a multitone complex is highest when its component tones are evenly spaced with respect to critical band and when the components are about equally loud. Loudness summation is independent of the way in which the sound is presented to the listener (e.g., binaural or monaural presentation) (Scharf, 1978).

TEMPORAL LOUDNESS SUMMATION AND SIGNAL DURATION

The perceived loudness of natural sounds varies also as a function of time. Temporal effects on loudness have been studied extensively, although there is considerable variability in results.

The relation between loudness and signal duration has been studied for both brief and long duration sounds. Such studies could aim, for example, at finding out how rapidly loudness reaches maximum value as duration of the sound is increased from a few milliseconds to hundreds of milliseconds. The studies into loudness adaptation and auditory fatigue use sounds of longer duration and are focused, for example, on how rapidly loudness decreases as the signal duration lengthens well beyond 1 s (Scharf, 1978).

Loudness summation occurs not only across frequency but also over time. Studies on the effect of signal duration and repetition rate of short pulses on loudness showed that loudness is integrated up to a 'critical duration' of about 100 ms, for example, (Buus, Florentine and Poulsen (1997), Verhey and Kollmeier (2002). The effect is usually referred to as temporal integration of loudness or temporal loudness summation.

According to Zwicker and Fastl (1999), the loudness of a tone burst decreases for durations smaller than about 100 ms. For longer durations, the loudness is almost independent of duration and remains constant. It decreases with shorter durations, and at a duration of about 10 ms is reduced by a factor of 2 (corresponding to a loudness level difference of 10 phon). However, as pointed out by Scharf (1978) and Moore (2003), precise values for temporal effects on loudness perception are not defined owing to contradictory experimental results.

THE BACKGROUND: PARTIAL MASKING

Perception of the loudness of the signal is dependent on whether it is presented to the listener in isolation or with another signal or signals as background. According to Scharf (1978), in isolation, a one-second long 1 kHz tone at 80 dB SPL has a loudness of 16 sones and is loud, but when heard against a 90 dB SPL white noise, it is perceived as soft. This reduction of loudness by a background noise is referred to as partial masking. Partial masking depends on the intensity of the masker as well as on its bandwidth and its frequency location relative to the frequency of the signal (Plack & Carlyon, 1995; Scharf, 1978).

STIMULI TYPE: COMPUTER GENERATED VS. REAL-LIFE SIGNALS

As pointed out by Skovenborg and Nielsen (2004), the results of subjective and objective loudness measurements are unavoidably influenced by the type of stimuli

traditionally used in psychoacoustics. Synthesised signals such as constant or pulsating sine waves, noise of varying bandwidth, and tone complexes are not representative of real life signals in terms of spectral content and dynamic variation. Traditional acoustic models derived from the analysis of relatively simple sounds may not be fully appropriate for the analysis of highly complex real life signals such as speech.

5.2.4 Factors influencing loudness perception: the listener

Certain listener-related factors need to be taken into consideration when designing psychoacoustic experiments and interpreting their results. While the experimenter can control such factors as binaural vs. monaural presentation of sounds or exposure to loud sounds prior to the actual listening tests, pathology of the auditory system lies outside the experimental control. Individual differences within the normal range as well as variations in decision making and personality that could affect the judgment of the sound by the listener, while also beyond the control of experimenter, need to be taken into account (Scharf, 1978).

A sound is perceived as being louder when presented in two ears than in one. According to Scharf (1978), pure tone or narrow band noise is about 1.7 times louder in two ears, on average. The phenomenon is sometimes referred to as binaural loudness summation (Plack & Carlyon, 1995; Scharf, 1978; Skovenborg & Nielsen, 2004).

The increase of binaural loudness relative to monaural loudness depends on the sound pressure level. According to Zwicker and Fastl (1999), for soft sounds (20 dB SPL), changing from monaural listening to binaural listening results in an increase in perceived loudness by a factor of about 2 (which corresponds to an increase in monaural sound pressure level of 8 dB). At high levels (80 dB SPL), the increase comes only to a factor of about 1.4 (corresponding to 6 dB SPL).

Perception of loudness can be influenced by auditory fatigue. Auditory fatigue caused by long-term exposure to loud sounds results in a temporary increase in the normal threshold accompanied by temporary loudness reduction. This reduction in loudness caused by fatiguing sound is similar to partial masking. Auditory fatigue is measured after the fatiguing stimulus has been removed. It can be quantified as the shift in a subject's absolute threshold at a particular frequency measured before and after the exposure to the fatiguing stimulus of a particular frequency and intensity. It has been found that the fatiguing sound reduces the loudness only of sounds whose intensity levels are much lower than its own: for loudness to be reduced, the judged tone has to be at least 40 to 50 dB SPL weaker than the fatiguing tone (Moore, 2003; Scharf, 1978).

According to (Scharf, 1978), perception of loudness depends on individual differences of the listeners, including variations in the auditory system or less specifiable variations in personality. Therefore, the best estimate of the slope of the loudness function can be obtained by averaging loudness judgements from a relatively large group of normal subjects.

Perception of loudness is affected by hearing impairment. For example, hearing impairment caused by defects in the cochlear is characterised by decreased sensitivity to the detection of soft sounds. However, with this type of impairment, loudness increases with intensity in an abnormally rapid way: a sound at a high intensity might sound equally loud to a hearing-impaired person as it does to a normally hearing listener. This phenomenon is known as recruitment (Plack & Carlyon, 1995).

5.2.5 Physiological correlates of loudness

As pointed out by Scharf (1978), rather than looking directly for physiological correlates of loudness, researchers try to determine how the auditory system codes sound pressure. 'Physiological events that are monotonic functions of sound pressure are likely correlates of loudness. Changes in sound pressure give rise to a sequence of mechanical events in the middle ear and cochlear that culminate in the bending of the 'cilia' of the hair cells. Just how the deformation of the cilia, which apparently triggers neural activity, varies with sound pressure is not known. It is known, however, that the maximum amplitude of displacement of the basilar membrane, on which the hair cells sit, is a linear or near linear function of sound pressure... Consequently, up to the point of transduction, the major physiological correlate of loudness is amplitude of displacement' (Scharf, 1978, p. 222).

The neural correlates of loudness are not understood very well. Scharf (1978) points at the possible correlation between the quantity of neural activity measured by the number of nerve impulses per second and perceived loudness: the more active the auditory nervous system is (that is, the higher the rate of firing in single neurons and the larger the number of active units involved), the greater the loudness. The depth or amplitude of displacement of the basilar membrane determines how fast the neurons fire, and the breadth or extent of the displacement determines how many of them fire. 'Widening the bandwidth of a sound, even without increasing overall intensity, broadens the area of displacement on the basilar membrane. Presumably, that is why loudness increases with stimulus bandwidth' (Scharf, 1978, p. 222).

According to Scharf (1978), although much of the variance in loudness seems to be related to the displacement pattern on the basilar membrane and the number of units

available for responding to the test signal, the notion of loudness as a simple correlate of total neural activity may be wrong or incomplete. He points at other possible approaches to the problem of loudness coding, such as, for example, a model of sensory magnitude proposed by Luce and Green (1972).

5.2.6 Loudness in speech studies

In speech research, perceived loudness has been studied primarily as a possible correlate of linguistic stress and syllable prominence, e.g., Beckman (1986), Campbell (1995), Crosswhite (2003), Fant (1997); Fry (1955); Gordeeva, Mennen and Scobbie (2003); Heldner (2001a, 2001b, 2003); Kochanski, Grabe, Coleman and Rosner (2005); Sluijter, Shattuck-Hufnagel, Stevens and Van Heuven (1995); Sluijter and van Heuven (1996); Sluijter, van Heuven and Pacilly (1997); Turk and Sawusch (1996). In these studies, various 'operationalisations' of perceived loudness have been proposed, such as measures of intensity and a variety of quantifications of the distribution of energy over the frequency spectrum. Perceived loudness has also been studied in relation to vocal effort (Brandt, Ruder, & Shipp Jr, 1969; Glave & Rietveld, 1975, 1979; Traunmüller & Eriksson, 2000).

Loudness variation forms important part of the prosodic signalling, along with duration and f_0. Compared to relatively robust measures of duration and f_0, intensity measures related to perceived loudness are sensitive to such factors as distance from the speaker to the microphone, and therefore less robust. More reliable loudness-related measures have been introduced, for example, measures of energy distribution over the frequency spectrum (spectral tilt).

Measures related to the spectral tilt have been discussed in Chapter 2, Section 2.1.3. Spectral tilt can be quantified in a number of ways, and one of the frequently used measures is H1*-A3* showing the difference (in dB) between the first harmonic H1 and the third formant spectral peak (A3), where asterisks show that the values are corrected for the boosting effect of the vocal tract transfer function. A related estimate of spectral tilt is the difference between the first and second harmonics H1*–H2* (Hanson, 1997; Hanson & Chuang, 1999; Sluijter et al., 1995; K. N. Stevens & Hanson, 1994).

Crosswhite (2003) reported on a study where spectral tilt was measured as the difference between loudness (in phon) and intensity (in dB). An increase in this measure was considered to indicate a less negative spectral tilt, leading to increased perceived loudness.

Sluijter and van Heuven (1996) studied spectral balance as an acoustic correlate of stress in Dutch. Spectral balance was determined by measuring the intensity in four non-overlapping adjacent frequency bands B1-B4: 0-0.5, 0.5-1.0, 1.0-2.0, and 2.0-4.0 kHz chosen so that the lowest band included the fundamental and the other bands included *F*1, *F*2, and *F*3 and *F*4, respectively. According to Heldner (2003) however, this measure should be regarded as reflecting spectral emphasis[25]. Heldner cites other studies in which spectral emphasis was measured as the difference between the overall intensity and the intensity in a low-pass-filtered signal, for example, Campbell (1995); Childers and Lee (1991); Traunmüller and Eriksson (2000).

In a perception experiment, Sluijter et al. (1997) manipulated syllable prominence by increasing intensity in the higher frequency bands. (The spectral balance of the syllables was therefore varied by increasing the levels of the frequency components above 500 Hz by 3, 6, or 9 dB, in either the initial or the final syllable). They hypothesised that varying the spectral balance rather than the overall intensity would result in a more realistic auditory cue of loudness for stress perception. Their results supported their hypothesis showing that increasing only the intensity level in the higher frequency bands provides a relatively strong stress cue, which could not be achieved by manipulation of the overall intensity level.

Kochanski et al. (2005) studied the acoustic correlates of stress in English. Their spectral slope estimate, S(t), is the slope of the best-fit to the Bark-binned spectrum between 500 Hz and 3000 Hz. Spectral tilt measures are usually found to be more robust correlates of perceived syllable prominence. However, Kochanski et al. (2005) did not find strong effects of spectral tilt on perception of prominence, compared to the results of Sluijter and van Heuven (1996), possibly due to such factors as different languages, a bigger and automatically processed speech corpus, and a different way to measure spectral tilt.

The overall intensity of the human voice increases with fundamental frequency, at least up to a mid-frequency of the speaker's f_0 range (Fant, Kruckenberg, & Liljencrants, 2000; Fant, Kruckenberg, Liljencrants et al., 2000). Conversely, a decrease in fundamental frequency is usually accompanied by a decrease in overall intensity. These adjustments are also related to the shape of the glottal pulse and spectral tilt.

[25] As discussed in Section 2.1.3, there is diversity in the literature as regards the terms denoting the slope of the spectrum (e.g., spectral tilt, spectral balance or spectral slope). Heldner (2003) distinguishes between 'spectral tilt' as explicitly representing the slope of the spectrum and 'spectral emphasis' as a measure of the relative contribution of the high-frequency components of the spectrum to the overall intensity.

Perceived loudness of speech sounds has been found to be related to the properties of the filter. For example, open vowels have been reported to have higher intrinsic intensity than close vowels (Honda, Maeda, & Sugito, 2007; Möbius, 2003).

A number of studies reported on the positive correlation between increased vocal effort and perceived loudness, e.g., Brandt et al. (1969), Glave and Rietveld (1975); Traunmüller and Eriksson (2000). Sluijter and van Heuven (1996) argued that an increase in vocal effort results in a more asymmetrical glottal pulse and a flatter spectral tilt (and thus increased syllable prominence). According to Klasmeyer (1997), objective measures of vocal effort include glottal pulse shapes or spectral damping of the harmonics in the frequency domain: steep pulses are usually a sign of high vocal effort. Traunmüller and Eriksson (2000) defined vocal effort as the quantity that ordinary speakers vary when they adapt their speech to the demands of an increased or decreased communication distance. (Another definition is found in Frick (1985), where vocal effort is characterised as a notion used to describe the intensity of pitch as it is produced by the speaker, as opposed to how loud it sounds to a listener, which might depend on the distance between the speaker and the listener.) Based on the assumption that the listeners are capable of distinguishing variations in a speaker's vocal effort from variations in their own distance from the speaker, Traunmüller and Eriksson (2000) measured vocal effort as the average rating, by a group of listeners, of the communication distance for each stimulus. The acoustic characteristics measured were segment durations, f_0, formant frequencies, SPL and spectral emphasis. Their results showed that, even though SPL is affected most by variations in vocal effort, it gives a listener no unambiguous cues and therefore cannot be used as an index of vocal effort. They concluded that more reliable information is conveyed by spectral emphasis.

In speech studies, estimation of perceived loudness is rarely performed using loudness models. Kochanski et al. (2005) in their study of acoustic properties of syllable prominence in English, used a loudness measure based on Stevens' loudness model rather than the more common RMS amplitude or peak amplitude measurements as it better approximates listeners' perception. However, they found no substantial difference between the loudness, RMS, and peak-to-peak classifier performance in terms of machine replication of the prominence judgements of human listeners.

Apart from the above, loudness, and in particular habitual loudness, has been an object of investigation in clinical speech studies. For example, decrease in habitual loudness has been found to be an important characteristic of a number of voice disorders.

In clinical studies, however, loudness is measured as intensity or sound pressure in dB SPL (Zraick, Marshall, Smith-Olinde, & Montague, 2004).

5.2.7 Loudness in vocal expression of emotion

In the research into vocal expression of emotion, loudness has been cited as a possible perceptual cue for emotion decoding. Prosodic features such as pitch, voice quality, loudness, and timing organisation of speech play a fundamental role in conveying emotions and attitudes in human communication. Specific emotions are communicated by particular combinations of pitch and loudness as well as the speaking rate. There are affects that are expressed loudly and others of which a low intensity is typical. For example, Scherer (1986, 2003) proposed acoustic profiles of emotional expressions which suggest that anger and happiness are signalled by increased pitch, increased loudness, and a faster rate of speech, whereas boredom and grief are characterised by low pitch and a slow speaking rate. As summarised in Frick (1985), contempt is loud and grief and boredom are soft. The experiments reported by Siegman and Boyle (Siegman & Boyle, 1993) proved that an increase in speech rate and loudness when speaking about fear and anxiety arousing events was associated with a corresponding increase in fear and anxiety. Similar correlation was found between sadness and depression and the decrease in speech rate and loudness. According to Scherer (2003), negative emotions and signs of aggression are characterised by increased speech intensity (which we interpret here as increased perceived loudness). Intensity and loudness are therefore often referred to as important factors to influence emotion encoding and recognition. However, Frick in his review of the role of prosodic features in communication of emotions (Frick, 1985) notes that these correlations are not always found and comments on the experiments with synthesised loudness that demonstrated relatively low contribution of loudness in communicating emotion. Frick's scepticism could be explained by the somewhat loose way he defines loudness in his review of the role of prosodic features such as pitch, loudness and speech rate in communicating emotions. For example, he treats intensity as well as subglottal pressure as analogous to loudness. Moreover, he uses the term 'stress' to mean 'loudness'. Similar treatment of perceived loudness as identical/synonymous to intensity is not infrequent in the literature, e.g., Bachorowski and Owren (2003).

Frick (1985) argues that f_0 and loudness (which he defines as stress and intensity) treated in isolation may not correspond to any auditory cues in perception of emotion. He suggests instead that the relevant feature for the communication of emotion should be vocal effort. He admits, however, that loudness may be related to the emotion dimension of

activation. Increased muscle tension that is a correlate of some emotions may lead to increase in the rate of respiration and faster speech. These various physiological responses lead to a correlation between activation and the prosodic features of increased loudness, raised pitch, increased pitch range, and increased rate of speech. Frick (1985) points out, however, that pitch alone is not sufficient to express an emotion; instead, the interaction between pitch and loudness in prosodic contours (patterns of pitch and loudness over time) needs to be considered.

Loudness-related acoustic measures such as mean intensity and energy distribution measures of one kind or another (related to spectral tilt, although not always referred to as such) are common in emotional speech research. For example, mean and standard deviation of voice intensity as well as high frequency energy in bands above 500 Hz are reported in Banse and Scherer (1996), Bänziger (2004), Juslin and Scherer (2005), Scherer (2003). Scherer (2003) points out, however, that more weight needs to be given to the study of what he refers to as 'distal cues' of the Brunswik lens model, such as perceived loudness, perceived pitch, perceived rhythm, etc. rather than basing the assumptions on affect communication and decoding exclusively on the objective speaker-related measures (such as intensity and f_0 variations). Schröder (2004b) comments that research suggests that emotions are not recognised on the basis of absolute intensity but rather on the basis of voice quality variations related to vocal effort of the speaker. He further hypothesises (p.49) that manipulating loudness of a synthesised stimulus while keeping voice quality constant should have a less prominent impact on the stimulus perception than varying the voice quality and keeping absolute loudness unchanged.

5.3 Loudness experiment 1

5.3.1 Introduction

As mentioned in the Introduction (Section 5.1), different voice qualities that were consistently associated with affects in the study reported in Chapter 4 as well as in earlier studies (e.g., Gobl & Ní Chasaide, 2003b) are characterised by different perceived loudness. To clarify to what extent the loudness differences among these voice qualities might influence the affective colouring that these voice qualities impart, two experiments were conducted, and the results are presented below.

The first experiment (Loudness experiment 1) tests the impact of loudness manipulations on the perception of affect using two types of synthesised stimuli: (1) stimuli incorporating distinct voice quality features including intrinsic loudness variations and (2) stimuli where voice quality (modal voice) was kept constant, but in which loudness was systematically modified. The experiment tests the following hypothesis: If loudness is the principal determinant in affect cueing for different voice qualities, then there should be little or no difference in the responses to the two sets of stimuli.

5.3.2 Material: synthesised stimuli

11 synthesised stimuli of a Swedish utterance 'ja adjö' ['jɑ a'jø] generated using the KLSYN88a formant synthesizer (Klatt, 1980) were used in the listening tests. The stimuli were of two types: (1) 'voice quality' stimuli and (2) 'loudness' stimuli. The 'voice quality' stimuli are differentiated in terms of voice quality characteristics and have loudness/intensity levels intrinsic to these voice qualities. In all of the 'loudness' stimuli voice quality (modal) was kept constant and the loudness was systematically manipulated to match the loudness of each of the original voice quality stimuli. The matching loudness of the 'loudness' stimuli was obtained in the course of a preliminary listening test (see Section 5.2.3 below).

'Voice quality' stimuli The synthesised voice qualities include modal voice, whispery voice, breathy voice, lax-creaky voice, harsh voice and tense voice. The description of the stimuli is given in Chapter 3. These stimuli simulate voice qualities according to the classification system in Laver (1980), with the exception of lax-creaky voice, which is conceptually an extension of the Laver framework.

'Loudness' stimuli Based on modal voice quality, five new stimuli were generated. Each of the new stimuli matched the loudness of one of the original non-modal voice

quality stimuli (whispery, breathy, lax-creaky, harsh and tense), e.g., there was a modal voice stimulus with the loudness matching that of tense voice, a modal voice stimulus with the loudness matching that of whispery voice, etc.

As discussed earlier in Section 5.2, loudness is defined as the subjective magnitude of a sound. Therefore, simple intensity normalisation was considered unsatisfactory in generating stimuli matching in loudness. Even though intensity manipulations were used as the method of generating the stimuli, the best loudness match had to be obtained in the course of a preliminary auditory experiment. The matching loudness stimulus would therefore not necessarily have the same sound intensity as the corresponding voice quality stimulus, but they would be perceived as equally loud.

5.3.3 Preliminary listening test: obtaining loudness matches

A preliminary listening test was carried out in order to find the stimuli that would best match in terms of perceived loudness each of the original voice quality stimuli. The test used stimuli with modal voice quality, but where the loudness was systematically changed.

The original modal voice quality was used as the basic stimulus. Its intensity level was increased/lowered in relatively fine steps of 1 dB to provide a selection of sample sounds, which could then be compared to the original stimuli in the course of auditory tests.

The procedure was similar to the loudness matching experiments common in psychoacoustic research. However, rather than letting the listeners regulate the gain control continuously to adjust the loudness of the test stimuli to match the reference stimulus, the listeners could choose from a set of discrete stimuli corresponding to relatively fine loudness steps.

A set of 24 stimuli was thus generated using the GoldWave v.4.26 software. Each stimulus was given a numeric value corresponding to the change in intensity in dB so that the 'quietest' stimulus (Stimulus –12) had an intensity level that was 12 dB less than that of the original modal voice stimulus and the 'loudest' stimulus (Stimulus +12) had an intensity level that was 12 dB higher than that of the modal voice. The original modal voice stimulus (Stimulus 0) was also included in the set. The total number of stimuli was 25.

To obtain the required intensity values, the amplitude of the original modal stimulus was multiplied by scaling factors corresponding to an increase/decrease of the intensity level by 1 dB. The resulting stimuli were arranged according to increasing intensity from the lowest intensity (Stimulus –12) to the highest intensity (Stimulus +12

dB), with the modal voice in the middle of the range. This order was kept constant as the range of stimuli was presented to the listeners. The listeners were informed of this arrangement of the stimuli prior to the experiment.

16 native speakers of Hiberno-English participated in the preliminary listening test. They were instructed to listen in turn to each of the five original voice quality stimuli (whispery, breathy, lax-creaky, harsh, and tense) which were called sound A, sound B, etc., and to select for each sound the best loudness match out of the range of 25 sample sounds. The participants were allowed to listen to the stimuli as many times as they needed to make a decision, and then to mark the responses on an answer sheet. Figure 5.4 illustrates the software interface used in the preliminary listening test.

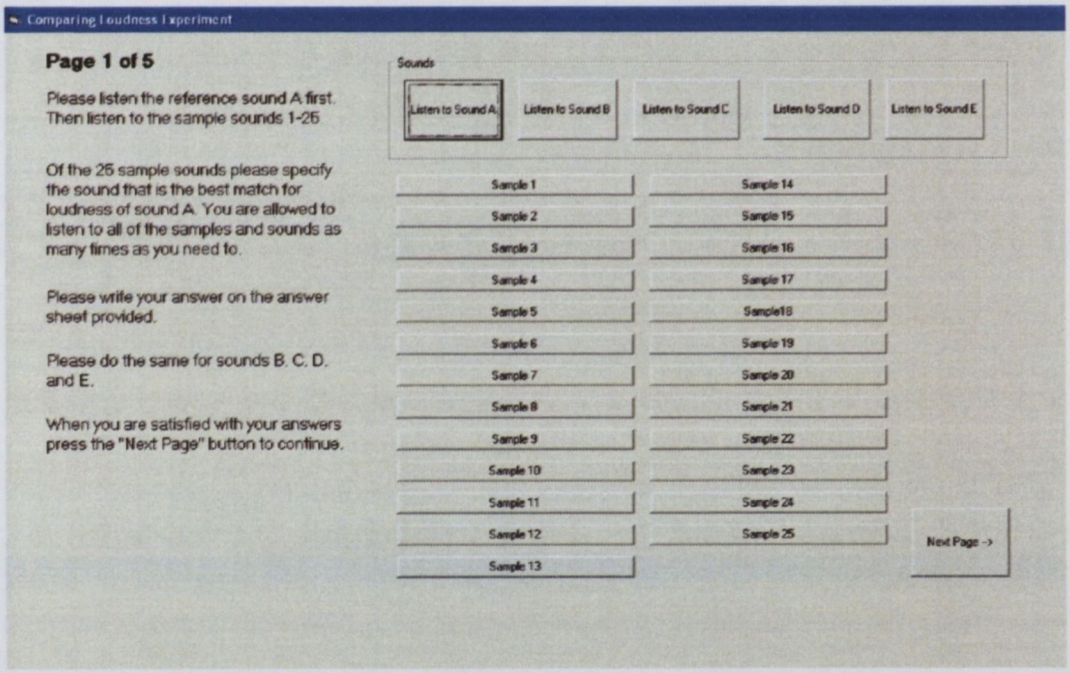


Figure 5.4. *Computer software interface used in the preliminary perception test.*

The stimuli were presented to the participants through a high quality speaker (Fostex 6301B) connected to a laptop, in a quiet room. The original voice quality stimuli (called sound A, sound B, etc.) were presented 5 times in a randomised order.

For each of the original voice qualities, the mean value was calculated, based on the best matching stimuli, obtained from the 16 participants' responses (a total of 16 x 5 = 80 responses). The average measure Intraclass Correlation Coefficient ICC (*R*) calculated to test the overall consistency of the participants in the ratings of the stimuli was found high at 0.99.

As the stimulus numbers corresponded to the dB change of the loudness level of the stimuli relative to modal, the mean values represent the required change to match loudness of the original voice qualities. These values are shown in Table 5.1, together with the corresponding scaling factors. The standard deviation of the mean values is represented in Table 5.1 in brackets. The same data is presented in Figure 5.5.

Five 'loudness' stimuli were generated by simply scaling the sample data of the original modal voice with each of the five scaling factors (see Table 5.1). These stimuli together with the original voice quality stimuli (11 in total) were used in the subsequent series of listening tests to investigate the role of loudness in the perception of affect.

Table 5.1. *Scaling factors and the difference in dB between the modal stimulus and the 'loudness' stimuli selected as best loudness matches for the original voice quality stimuli.*

Stimuli	Scaling factor	Difference in dB and SD
m_whispery	0.433	-7.27 (1.17)
m_breathy	0.628	-4.04 (1.34)
m_lax-creaky	0.727	-2.77 (0.86)
m_harsh	1.352	+2.62 (1.47)
m_tense	1.427	+3.09 (1.18)

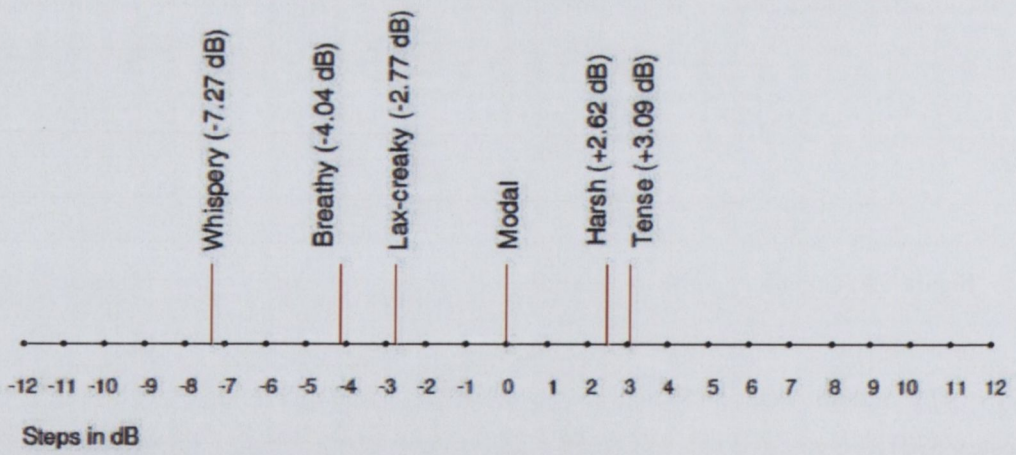


Figure 5.5. *The 'loudness stimuli' on the dB scale relative to modal voice. Values in brackets show the difference in dB between the overall intensity of the 'loudness' stimuli and modal voice.*

5.3.4 Listening tests

METHOD

The 11 stimuli (breathy; modal with breathy loudness; whispery; modal with whispery loudness; lax-creaky; modal with lax-creaky loudness; tense; modal with tense loudness; harsh; modal with harsh loudness; and modal) were presented to 16 subjects, native speakers of Hiberno-English, females.

The perception test was conducted according to the procedure described in Gobl and Ní Chasaide (2003b) and in experiments described in Chapter 4 of this thesis as a series of six subtests. In each subtest, 10 randomisations of 11 stimuli were presented to the participants, and responses were obtained for a pair of opposite affective attributes. The pairs of affective attributes tested were *apologetic-indignant, bored-interested, intimate-formal, relaxed-stressed, sad-happy* and *scared-fearless*. The participants were asked to judge for each stimulus whether the speaker sounded more sad or happy, etc., and to mark their response on the answer sheet. The ratings were interpreted as a seven point scale ranging from -3 to + 3, where 0 corresponded to no affect perceived, and plus or minus 1, 2 or 3 to mild, moderate and strong presence of an affect respectively. The experimental procedure is described in detail in Chapter 3.

RESEARCH QUESTIONS

- To what extent the 'Loudness' stimuli (in which loudness variations intrinsic to the original voice quality stimuli is kept but in all of which voice quality is kept constant – modal) are capable of signalling affect?
- Do 'Loudness' stimuli yield the same affective ratings as the 'voice quality' stimuli? If not, what is the difference in affective ratings?

DESIGN

To compare the effect of loudness voice quality on the strength of affective ratings, a 5 x 2 factorial design was used in the statistical analysis, with 2 within-subjects factors: 'Loudness' (5 levels: loudness of whispery, breathy, lax-creaky, harsh, tense) and 'Voice' (2 levels: non-modal voice quality with a concomitant loudness, and modal voice quality). The dependent variable was the mean rating score for each stimulus averaged across 10 randomisations for each subject.

A two-way repeated measures ANOVA was done for each of the six subtests separately. The alpha level was set to 0.05. The Mauchley test indicated that the data did

not meet the assumption of sphericity, and therefore a Greenhouse-Geisser correction was applied to the degrees of freedom in the analysis. Bonferroni corrections were further applied to account for multiple comparisons in the post-hoc tests. The detailed output of ANOVA and the corresponding post-hoc tests are documented in Appendix 9.

As the factorial design did not include modal voice, a series of simple contrasts was conducted in a separate test in which modal voice was compared to the other 10 stimuli. The output of ANOVA is documented in Appendix 9.

RATERS' AGREEMENT

Raters' agreement was calculated as ICC (*r, R*) as described in Chapter 3 and is reported below for each subtest. Percent agreement data for each stimulus in each subtest is given in Appendix 10. Similar to the procedure described in Chapter 3, Pearson's correlation of the mean ratings of individual raters and the group mean for each subtest was calculated. This allows one to establish whether there were listeners who rated the stimuli very differently from the group as a whole. The data are presented in Appendix 11. The scatter plots in Appendix 12 in which the mean ratings are plotted against the standard deviation for each individual rater serve as an indication of the listener-judges' confidence in ratings: the standard deviation of 0-1 would be interpreted as suggestive of reasonable confidence in ratings. The same data can be inferred from the interquartile range of the box plots in Appendix 8.

5.3.5 Results and discussion

The results are presented in Figure 5.6. The ratings of the 'voice quality' stimuli are shown in red, the ratings of the corresponding 'loudness' stimuli are shown in black. Asterisks show statistically significant difference between the two series of stimuli. The results will be discussed separately for each of the subtests, and then summarised. Similar to the cross-language study reported in Chapter 4, the discussion will be focused primarily on the affective ratings above the arbitrarily chosen threshold of 1 (described in Chapter 3).

Apologetic-Indignant

Significant main effects were found for Loudness, $F(2.08, 31.2) = 80.62$, $p < 0.001$, $\eta_p^2 = 0.84$, and Voice x Loudness interaction $F(1.85, 27.73) = 23.11$, $p < 0.001$, $\eta_p^2 = 0.61$. The Voice effect was not significant, $F(1, 15) = 1.59$, $p = 0.22$, $\eta_p^2 = 0.1$. This suggests that the results of the manipulation that involved replacing a non-modal voice quality with modal while keeping the perceived loudness constant depend on the actual voice quality and its intrinsic loudness. As evident from Figure 5.6, for all voice qualities, replacing a non-modal voice quality with modal resulted in a statistically significant decrease in ratings (except lax-creaky voice, where statistical significance was not reached). None of the 'loudness' stimuli yielded the ratings above the threshold. Interestingly, whispery voice and breathy voice yield relatively high ratings for *apologetic*, but lax-creaky voice does not.

The results of the raters' agreement were as follows: ICC (r) = 0.73, ICC (R) = 0.98. This can be interpreted as substantial raters' agreement.

Bored-Interested

Significant main effects were found for Loudness, $F(1.93, 29) = 70.9$, $p < 0.001$, $\eta_p^2 = 0.82$, Voice, $F(1, 15) = 29.1$, $p < 0.001$, $\eta_p^2 = 0.66$ and Voice x Loudness interaction $F(1.81, 27.2) = 16.65$, $p < 0.001$, $\eta_p^2 = 0.53$. This suggests that as in the previous test, the results of the manipulation that involved replacing a non-modal voice quality with modal while keeping the perceived loudness constant depend on the actual non-modal voice quality and its intrinsic loudness. For breathy, whispery and in particular lax-creaky voice this resulted in a significant decrease in ratings for *bored*, while for harsh and tense voice there was no such effect. Lax-creaky and (weakly) whispery voice were found to cue *bored*. Both tense and m_tense stimuli were associated with *interested*.

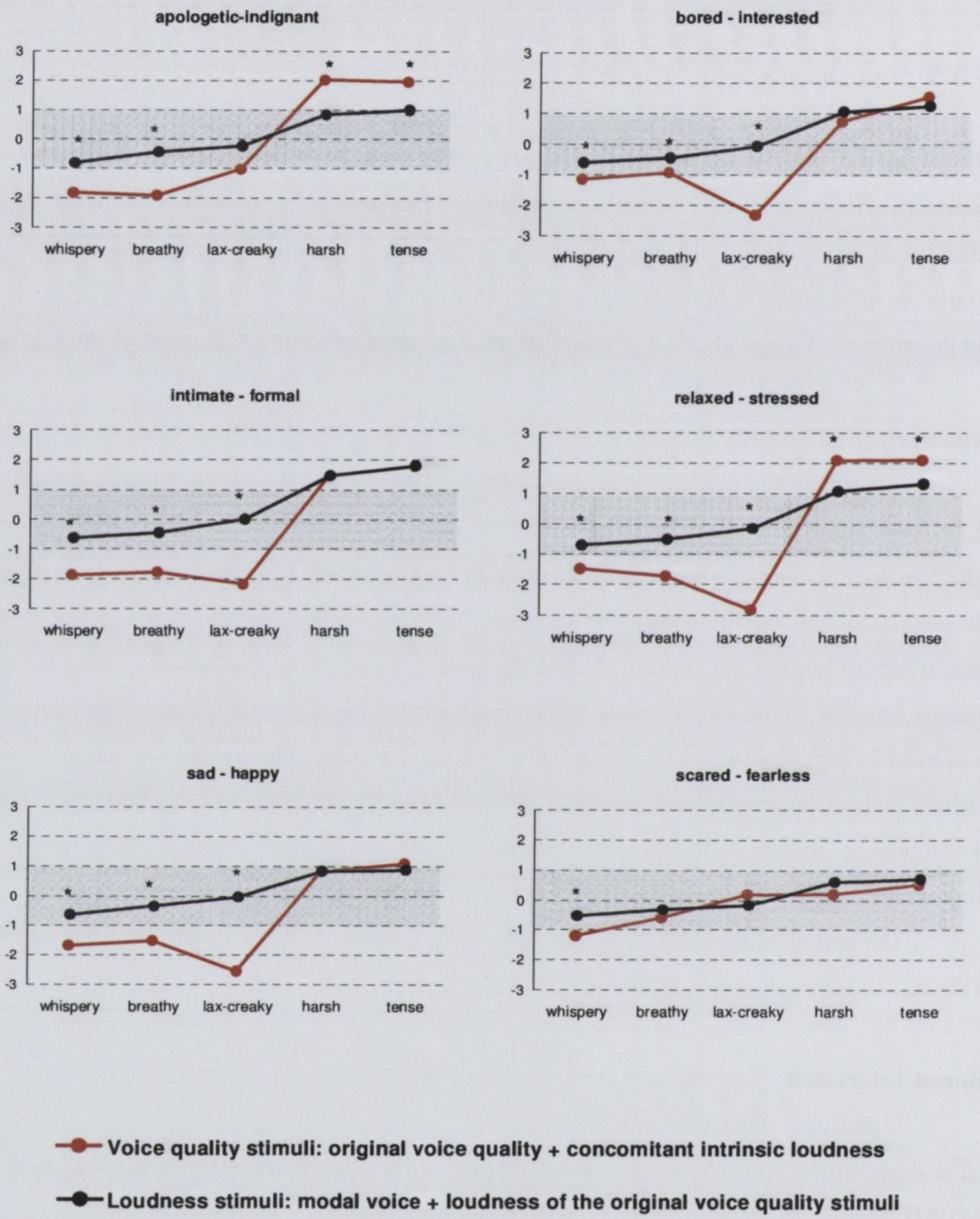


Figure 5.6. *Effects of the loudness and voice quality on affect ratings. Plotted are estimated marginal means. Ratings for 'voice quality' stimuli (original VQ loudness + original VQ) are shown in red; ratings for the 'loudness' stimuli (original VQ loudness + modal voice) are shown in black. Asterisks show significant difference in ratings between the two series of stimuli at $p < 0.05$. Shaded is the area of weak affect cueing.*

The raters were coherent in rating the stimuli in this subtest, as shown by the ICC (r) = 0.74, ICC (R) = 0.98. The measures suggest substantial raters' agreement.

Intimate-Formal

Significant main effects were found for Loudness, $F(2.54, 38{,}14) = 199.1$, $p < 0.001$, $\eta_p^2 = 0.93$, Voice, $F(1, 15) = 53.73$, $p < 0.001$, $\eta_p^2 = 0.78$ and Voice x Loudness interaction $F(2.04, 30.66) = 19.37$, $p < 0.001$, $\eta_p^2 = 0.56$. As in the tests above, the difference in ratings between the 'loudness' stimuli and 'voice quality' stimuli appears to depend on the intrinsic loudness of a particular non-modal voice. Overall, replacing whispery, breathy, lax-creaky voice with modal while keeping the intrinsic loudness differences of these non-modal voice qualities resulted in significant decrease in ratings for the 'loudness' stimuli: all are below the threshold. All these non-modal voice qualities cue *intimate* (to support findings in Chapter 4 for Hiberno-English listeners). Here loudness alone appears to be insufficient for affect cueing and voice quality is crucial. On the other hand, for harsh and tense voice no effect of voice quality was found. Both non-modal ('voice quality') stimuli and modal ('loudness') stimuli yield high ratings for *formal*. Here changing voice quality to modal does not have impact on affective association and loudness alone works as well.

The raters were very consistent in rating the stimuli in this subtest: ICC (r) = 0.86, ICC (R) = 0.99. This suggests outstanding raters' agreement.

Relaxed-Stressed

Significant main effects were found for Loudness, $F(1.52, 22.78) = 218.16$, $p < 0.001$, $\eta_p^2 = 0.94$, Voice, $F(1, 15) = 21.9$, $p < 0.001$, $\eta_p^2 = 0.59$ and Voice x Loudness interaction $F(1.91, 28.64) = 84.39$, $p < 0.001$, $\eta_p^2 = 0.85$. As in the tests above, the difference in ratings between the 'loudness' stimuli and the 'voice quality' stimuli depends on the loudness of a particular non-modal voice. The 'loudness' stimuli that have the loudness of whispery, breathy and lax-creaky voice yield significantly lower ratings than their counterparts from the 'voice quality' stimuli series. Their ratings are all below threshold, while the corresponding non-modal voice qualities all successfully cue *relaxed* (this association also supports the findings in Chapter 4 for Hiberno-English listeners). Here again loudness alone appears to be ineffective and voice quality is needed. For tense and harsh voice the pattern is similar to the one for the *indignant* affect: changing the voice quality to modal while keeping the loudness of tense or harsh voice results in a decrease in

ratings. Note, however, that the 'loudness' stimuli do yield moderate ratings for *stressed*, although harsh and tense voice qualities get significantly higher ratings.

In this subtest, the raters' agreement was also very high (outstanding): ICC (r) = 0.89, ICC (R) = 0.99.

Sad-Happy

Significant main effects were found for Loudness, $F(1.69, 25.33) = 69.32$, $p < 0.001$, $\eta_p^2 = 0.82$, Voice, $F(1, 15) = 72.79$, $p < 0.001$, $\eta_p^2 = 0.83$ and Voice x Loudness interaction $F(2.18, 32.64) = 35.17$, $p < 0.001$, $\eta_p^2 = 0.7$. Therefore, the difference in ratings between the 'loudness' stimuli and the 'voice quality' stimuli depend on the loudness of a particular non-modal voice. Similar to the results of the *intimate-formal* and *relaxed-stressed* subtests, replacing whispery, breathy, lax-creaky voice with modal while keeping their loudness resulted in significantly lower affective ratings, as suggested by the comparison of the ratings of these voice qualities and the corresponding 'loudness' stimuli. None of the 'loudness' stimuli were rated highly for *sad*. The non-modal voice qualities all successfully cue *sad,* with lax-creaky yielding the highest ratings. (This also supports the findings for Hiberno-English listeners in Chapter 4.) Here again lowering the loudness level alone is not sufficient for affect cueing and voice quality is required. The effect of stimulus manipulation for harsh and tense voice is similar to that in the *bored-interested* subtest: harsh and tense voice in the 'voice quality' stimuli series and the corresponding stimuli from the 'loudness' stimuli series yielded near-threshold ratings for *happy*.

The raters' agreement was substantial at ICC (r) = 0.79, ICC (R) = 0.98, suggesting that the listeners were uniform in rating voice to affect associations.

Scared-Fearless

Significant main effect was found for Loudness, $F(1.28, 19.15) = 4.25$, $p = 0.004$, $\eta_p^2 = 0.22$. Effects for Voice $F(1, 15) = 2.24$, $p = 0.15$, $\eta_p^2 = 0.13$.and Voice x Loudness interaction $F(1.77, 26.6) = 1.4$, $p = 0.25$, $\eta_p^2 = 0.09$, were not significant. In this subtest, the stimuli, irrespective of the loudness and matched voice quality largely failed to cue affect. The exception is whispery voice, which was associated with *scared,* although the rating was quite low. (Note that in the experiments reported in Chapter 4 it did not yield high ratings for *scared* either).

In this test, the raters' agreement was poor as indicated by ICC (r) = 0.14, ICC (R) = 0.72. The results are in keeping with those reported in Chapter 4, where the raters' agreement on the *scared-fearless* scale was generally low. This poor raters' agreement is

also evident in quite random scatter of responses for this subtest in Appendix 12 as well as in the percent agreement data in Appendix 10.

In summary, for five tests out of six, with the exception of *scared-fearless,* the Voice x Loudness interaction was found significant. This indicates that the difference in ratings between the 'voice quality' stimuli and the corresponding 'loudness' stimuli depends both on the intrinsic loudness and voice quality characteristics of a particular quality. In other words, the impact of replacing a non-modal voice quality with modal while keeping its original loudness on the strength of affective ratings depends on both the intrinsic loudness and voice quality characteristics of a particular stimulus.

The trend in Figure 5.6 is very similar for the *bored-interested, intimate-formal, sad-happy* subtests. In these subtests, the ratings of whispery, breathy, lax-creaky voice are significantly higher than the ratings of their counterparts from the 'loudness' stimuli series. Tense and harsh voice, on the other hand, yielded the same ratings as their 'loudness' stimuli counterparts. For two affects, *indignant* and *stressed*, the effect of loudness matching was more pronounced for tense and harsh voice. Here, tense and harsh voice yielded higher ratings as cueing *indignant* and *stressed* than modal voice with the loudness of tense and harsh voice.

Replacing any specific non-modal voice quality with modal voice while keeping the same loudness lowers the affective ratings particularly markedly for lax-creaky voice as opposed to its counterpart from the 'loudness' series for such affects as *sad, bored, intimate* and *relaxed.* Note that the magnitude of effect is less prominent for breathy and whispery voice. Note also that for *bored, intimate, relaxed* and *sad* lax-creaky voice yields significantly higher ratings than whispery or breathy voice.

In one subtest, *scared-fearless*, none of the stimuli achieved high affective rating, and there was no difference in the ratings obtained by the 'loudness' and the 'voice quality' stimuli. One exception here is whispery voice that (weakly) cues *scared* and that got ratings significantly higher than those of its 'loudness' counterpart.

Maximum mean ratings

Figure 5.7 shows the mean rating for the stimulus that received the highest rating (maximum mean rating) for each affect for each stimulus type. Table 5.2 shows the stimuli yielding the highest rating for the two stimulus types, 'voice quality' stimuli and 'loudness' stimuli for each of the affects tested. As is clear from Figure 5.7, 'voice quality' stimuli and 'loudness' stimuli perform rather differently in terms of affect signalling: in the

majority of cases, the 'voice quality' stimuli generate higher affective ratings than the 'loudness' stimuli. For eight affects out of 12, the ratings yielded by the 'voice quality' stimuli were markedly higher than the 'loudness' stimuli. In the remaining four affects, *formal, interested, happy* and *fearless*, differences were very small and not statistically significant. For *happy* and *fearless,* the ratings were low regardless of stimulus type.

In the 'loudness' series, only for 4 out of 12 affects (*formal, stressed, interested* and *indignant*) was the maximum mean rating 1 or higher. In contrast, maximum mean ratings for the 'voice quality' stimuli were above 1 with a single exception, *fearless*. These results support the hypothesis that the contribution of loudness is not the main determinant of the affective colouring conveyed by different voice qualities. The hypothesis is supported mainly for 'lax' voices (breathy, whispery, lax-creaky) for all affects tested. '

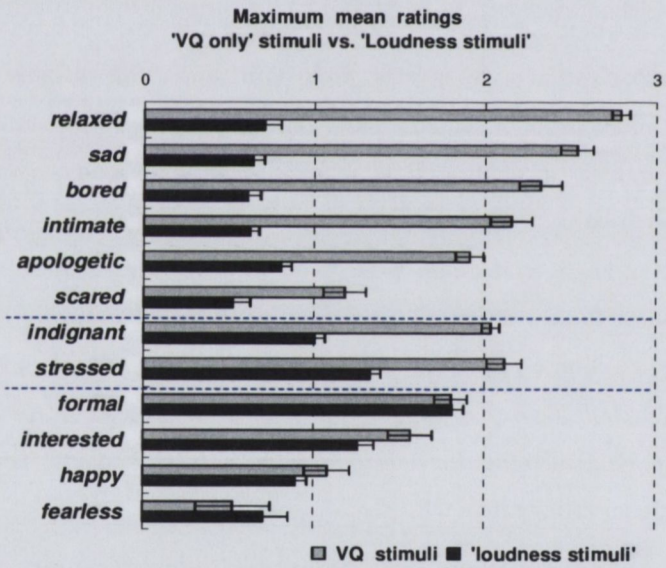


Figure 5.7. *Maximum mean rating and estimated standard error of the mean. Affect rating: 3 = max; 0 = none.*

Table 5.2. *Stimuli yielding the highest mean rating for each affect.*

Affect	**'Voice quality' stimuli**	**'Loudness' stimuli**
relaxed, sad, bored, intimate	lax-creaky	m_whispery
apologetic	breathy	m_whispery
scared	whispery	m_whispery
indignant	harsh	m_tense
stressed, formal, interested, happy, fearless	tense	m_tense

This does not necessarily mean that the role of loudness is non-existent. For certain affects, particularly *formal*, and to a somewhat lesser extent *stressed* and *interested* loudness alone results in a fairly high rating. This suggests that for these particular affects loudness acts as an important perceptual cue. It is worth noting that for *formal* the addition of tense voice quality did not yield higher affective rating, as modal voice with the loudness of tense and tense voice proper got ratings that were not significantly different.

Table 5.2 specifies which of the stimuli in either series, 'voice quality' stimuli or 'loudness' stimuli, was most effective in cueing the individual affective states. In the case of the 'voice quality' series, each of the voice qualities yielded highest ratings for at least one affective state. In the case of the 'loudness stimuli', the highest mean ratings were always associated with one of the two stimuli: the m_whispery and the m_tense. As can be seen from Table 5.1 and Figure 5.5, these two stimuli represent the extreme ends of the loudness continuum, with m_whispery being the quietest and m_tense being the loudest.

In the case of m_whispery, it is the most highly rated for the affects *relaxed, sad, bored, intimate, apologetic* and *scared,* i.e. the top six affects shown in Figure 5.7. However, in each of these cases the actual ratings were very low. For these affects, the voice quality appears to be of crucial importance and the difference between the loudness and voice quality stimuli is very pronounced.

In the case of m_tense stimulus (modal voice with the loudness of the original tense voice), the loudness level as such does appear to have some affect cueing function, particularly for *formal, stressed* and *interested*. For these affects, the contrast between the two types of stimulus is less prominent. It is worth noting that the m_tense stimulus is in all cases associated with the high activation affects. It would seem therefore that loudness as such does facilitate the perception of high activation. The opposite is not true: m_whispery (the quietest stimulus) does not seem to contribute much to the perception of low activation states, where voice quality appears to be crucial.

The results of this loudness experiment are in keeping with those obtained for Hiberno-English speakers reported in Chapter 4 (and also in Yanushevskaya, Gobl and Ní Chasaide (2005) and support the broad findings in terms of both voice-to-affect mapping and the rating levels yielded by the voice quality stimuli. Table 5.3 gives a brief comparison of results.

Table 5.3. *Voice quality-to-affect association by Hiberno-English listeners in Yanushevskaya et al. (2005) and present experiment; max mean ratings >1 are shown in bold type.*

Voice quality	Affects associated with voice quality	
	Earlier experiment	**Present experiment**
whispery	scared, **intimate, apologetic**	**scared**
breathy	-	**apologetic**
lax-creaky	**relaxed, sad, bored**	**intimate, relaxed, sad, bored**
harsh	[not used]	**indignant**
tense	**indignant, formal, fearless, stressed**, interested, happy	**formal**, fearless, **stressed, interested, happy**

5.3.6 Conclusions

The analysis of results shows that the pattern is different for the breathy, whispery and lax-creaky voice on the one hand and for the harsh and tense voice on the other. The 'loudness' stimuli counterparts of breathy, whispery and lax-creaky voice yield significantly lower affective ratings than the original non-modal stimuli. Changing the original voice quality to modal for tense and harsh voice leads to less pronounced differences in affective ratings, in particular in ratings for *interested, formal* and *happy*.

Overall, the results of the Loudness experiment 1 show that variation in voice quality is essential to the expression of affect, while loudness variation on its own is ineffective for affect cueing. Stimuli incorporating voice quality variations consistently obtained higher ratings than the 'loudness' stimuli.

Loudness may play a role in the cueing of some affects such as *formal*, *stressed* and *interested*, but overall, loudness on its own appears to be an insufficient cue to affect – even the quietest modal voice fails to convey low activation affects adequately.

5.4 Loudness experiment 2

5.4.1 Introduction

Loudness experiment 1 demonstrated that changing loudness while maintaining constant voice quality is rather ineffective in cueing affect, in particular for low activation states. However, this does not necessarily mean that the intrinsic loudness differences which tend to be correlated with particular voice qualities are irrelevant. Given that in human speech production there is a natural tendency towards co-variation of voice quality and loudness, it can be hypothesised that loudness differences do play an important role – but only when this loudness variations occur with the appropriate voice quality. Loudness experiment 2 reported below tests the hypothesis that equalising the perceived loudness of different voice quality stimuli will have relatively little impact on affective ratings.

5.4.2 Material: synthesised stimuli

The construction of the new set of stimuli for perception tests of Loudness experiment 2 consisted of two steps. The first step involved an increase or a decrease of the loudness of all the original non-modal voice quality stimuli (whispery voice, breathy voice, lax-creaky voice, harsh voice and tense voice) to match the loudness of the modal voice stimulus. In the second step, for each of the new loudness normalised voice quality stimuli a 'louder' version and a 'quieter' version were generated. Thus, for example, there was a stimulus with whispery voice quality and the loudness of modal voice and its 'louder' and 'quieter' versions. The difference between the 'loudness' versions was arbitrarily set to plus or minus 2 dB (see Figure 5.8), i.e. the stimulus having the loudness of modal voice was scaled by a factor of 1.26 for the 'louder' version and by a factor of 0.79 for the 'quieter' version.

Step 1 To generate the new voice quality stimuli with the loudness matching that of the modal stimulus, the amplitude of each of the original voice quality signal was multiplied by the reciprocal of the corresponding scaling factors used in Loudness experiment 1 (see Table 5.1) resulting in a uniform increase or decrease in overall amplitude. For example, the scaling factor needed to generate the modal voice quality stimulus with the loudness matching that of the tense voice was 1.427, which corresponds to an increase in intensity level by 3.09 dB (20 log 1.427). Therefore, to generate a tense

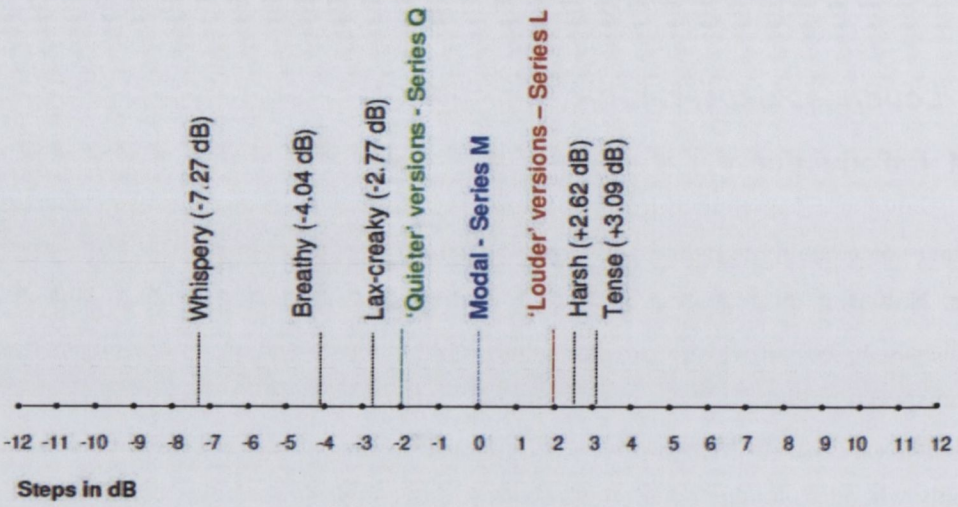


Figure 5.8. *The original voice quality stimuli relative to modal voice and the 'quieter' and 'louder' versions used in Loudness experiment 2.*

voice stimulus with the loudness matching that of the modal voice, the amplitude of the original tense voice stimuli needs to be multiplied by the reciprocal of this scaling factor (1/1.427 = 0.701), which would correspond to the decrease in intensity level by 3.09 dB. Similarly, the best loudness match for lax-creaky voice was modal voice multiplied by a factor of 0.727, which corresponds to a decrease of 2.77 dB. To get a lax-creaky voice with the loudness of modal, an increase of 2.77 dB should be carried out on the original lax-creaky voice, and so the multiplication factor of 1.376 was applied (1/0.727 = 1.376). The scaling factors used to synthesise the voice quality stimuli with equal loudness for experiment 2 are presented in Table 5.4. The software used for stimuli generation was GoldWave v.4.26. The resulting stimuli formed Series M.

Table 5.4. *The scaling factors used in the generation of loudness stimuli (Loudness experiment 2). The letter 'M' after the name of the stimulus in the table indicates that the loudness of the original voice quality stimulus was set to that of modal.*

Series M stimuli	Scaling factor
whispery_M	2.309
breathy_M	1.592
lax-creaky_M	1.376
modal	1
tense_M	0.701

Step 2 From Series M stimuli, two more stimuli series were generated: Series L ('louder versions') in which the intensity level of all stimuli was increased by 2 dB and

Series Q ('quieter versions') in which the intensity level of all stimuli was reduced by 2 dB. As mentioned, the particular choice of intensity increase/decrease was arbitrary. As the aim was to compare the perception of stimuli differing in voice quality characteristics but having the same loudness as well as to test the effect of any perceivable difference in loudness, a 2 dB difference between the intensity level of the stimuli in the three groups was considered sufficient.

As is clear from Figure 5.8, neither the 'louder' (+2 dB) nor the 'quieter' (-2 dB) versions of the new stimuli would match the loudness of any of the original voice quality stimuli. For example, the intensity level of the new tense_L stimulus (+2 dB) was 0.62 dB lower than that of the original tense voice quality.

The resulting number of stimuli was 6 x 3 = 18. Given that the results obtained for harsh voice in the course of the previous experiment (reported in Section 5.2 above) were similar to those of tense voice as well as for practical reasons the harsh voice quality stimulus was excluded from the list of stimuli used in Loudness experiment 2. Thus, 15 stimuli in total were used. Table 5.5 represents the complete set of stimuli and their file names.

Table 5.5. *The stimuli used in Loudness experiment 2.*

Series Q: (minus 2 dB)	**Series M (modal loudness)**	**Series L: (plus 2 dB)**
whispery_Q	whispery_M	whispery_L
breathy_Q	breathy_M	breathy_L
lax-creaky_Q	lax-creaky_M	lax-creaky_L
tense_Q	tense_M	tense_L
modal_Q	modal	modal_L

5.4.3 Listening tests

METHOD

The 15 stimuli were randomised 10 times and presented to the participants in a series of six sub-tests; the same pairs of affective attributes were used in the listening tests as in the previous experiment and in the experiments described in Chapter 4: *apologetic-indignant, bored-interested, intimate-formal, relaxed-stressed, sad-happy* and *scared-fearless.* The participants in the experiment were 16 native speakers of Hiberno-English, females. The stimuli were presented to the participants over a high quality speaker in a quiet room. Refer to Chapter 3 for further details regarding the experimental procedure.

RESEARCH QUESTIONS

- How do stimuli differing in voice quality but with loudness normalised yield affective ratings?
- To what extent does loudness variation contribute to affect signalling when stimuli differ in voice quality?

DESIGN

To assess the extent to which equalising the loudness of different voice quality stimuli might have on affective ratings, a 5 x 3 factorial design was used in the statistical analysis. The two within-subject factors were 'Voice' (five levels: whispery, breathy, lax-creaky, modal, tense) and 'Loudness' (three levels: 'Q' ('quieter' version, -2 dB), 'M' (modal loudness) and 'L' ('louder' version, +2 dB). The dependent variable was the mean rating score for each stimulus averaged across 10 randomisations for each subject. The two-way repeated measures ANOVA was performed for each of the six subtests separately. The alpha level was set to 0.05. As the data did not meet the sphericity assumptions as indicated by the Mauchley test in most cases, a Greenhouse-Geisser correction was applied to the degrees of freedom in the analysis. The output of ANOVA and the corresponding post-hoc tests are documented in Appendix 14.

The raters' agreement was measured as outlined in Chapter 3 and in Loudness experiment 1 above. ICC (*r, R*) was computed to assess the coherence of the stimuli ratings. Percent agreement was calculated to assess the agreement in rating individual stimuli. The rating of each individual listener was correlated to the group ratings as another measure of coherence using Pearson's *r*. The agreement data are given in Appendices 15 - 16. The listeners' agreement in ratings a particular stimulus can be further inferred from the interquartile ranges in the box plots in Appendix 13. The data on the listeners' confidence in ratings can be inferred from the scatter plots in Appendix 17 where for each individual rater mean rating value across 10 randomisations is plotted against the standard deviation.

5.4.4 Results

The results are shown in figure 5.9 and will be discussed for each subtest separately, and subsequently summarised.

Apologetic-Indignant

Significant main effects were found for Voice, $F(2.36, 35.35) = 75.5$, $p < 0.001$, $\eta_p^2 = 0.83$, and Loudness $F(1.27, 19.08) = 22.21$, $p < 0.001$, $\eta_p^2 = 0.6$. The two-way interaction

was also found to be significant, $F(5.64, 21.97) = 3.85$, $p = 0.006$, $\eta_p^2 = 0.2$. This suggests that loudness manipulation has different consequences for different voice qualities: increasing the loudness increases the ratings of modal and tense voice and has the opposite effect on the ratings of whispery, breathy and lax-creaky voice. As evident from Figure 5.9, loudness normalisation of the non-modal voice quality stimuli to that of modal voice does not impede affective ratings: all non-modal stimuli from Series M yield above threshold affective ratings.

The listeners showed substantial agreement in ratings in this subtest, as suggested by ICC (r) = 0.78, ICC (R) = 0.98.

Bored-Interested.

Significant main effects were found for Voice $F(1.64, 24.64) = 19,65$, $p < 0.001$, $\eta_p^2 = 0.57$; Loudness $F(1.22, 18,3) = 44.22$, $p < 0.001$, $\eta_p^2 = 0.75$, and Voice x Loudness interaction $F(5.5, 82.51) = 4.87$, $p < 0.001$. $\eta_p^2 = 0.25$. Loudness manipulation has therefore similar effect here as in the test above. In this test, however, increasing the loudness results in an increase in the rating for *interested* for modal and tense voice, and also for whispery and breathy voice. The results for lax-creaky voice are clearly separate from those of breathy and whispery voice: increasing its loudness lowers its ratings for *bored*, but all its loudness variants are rated on the *bored* end of the scale. The effect of loudness normalisation to modal is different from the above test: only tense and lax-creaky voice stimuli from Series M appear to cue affect.

The raters' agreement in this test was lower than in the previous one and can be interpreted as moderate: ICC (r) = 0.55, ICC (R) = 0.95.

Intimate-Formal

Significant main effects were found for Voice $F(2.48, 37.21) = 76.23$, $p < 0.001$, $\eta_p^2 = 0.84$; Loudness $F(1.16, 17.2) = 15.98$, $p < 0.001$, $\eta_p^2 = 0.52$, and the two-way interaction $F(4.19, 62.8) = 3.86$, $p = 0.006$, $\eta_p^2 = 0.21$.

Loudness manipulation has the same effect here as in the tests above for modal and tense voice: as the loudness increases, the ratings for *interested* increase.

The ratings of lax-creaky, breathy and whispery voice from Series Q are not significantly different from the corresponding stimuli from Series M. Increasing their loudness further (Series L) has no effect on the ratings for *intimate* yielded by lax-creaky voice and results in lowering the ratings of breathy and whispery voice.

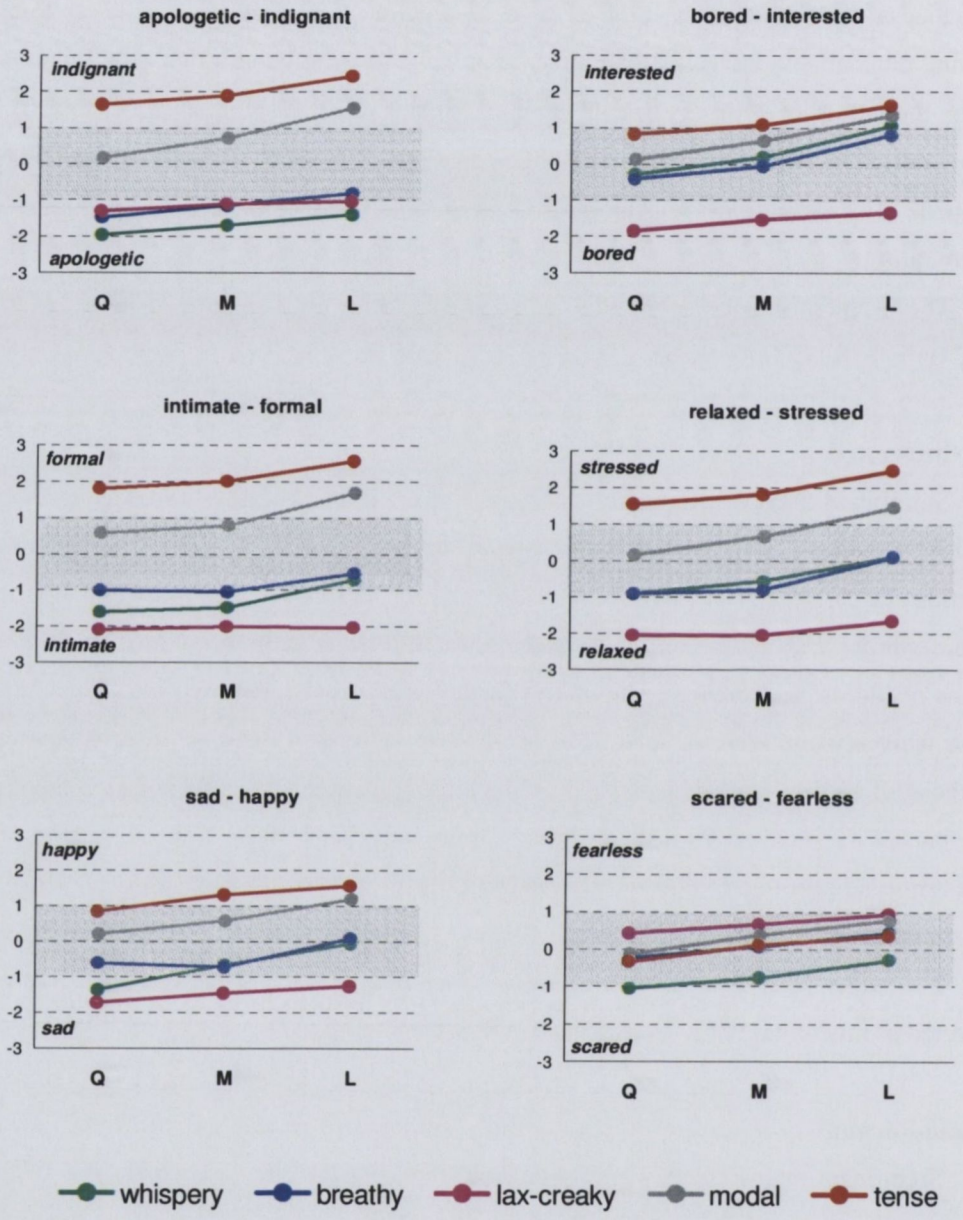


Figure 5.9. *Effects of the loudness manipulation in different voice qualities on affective ratings. Plotted are estimated marginal means. Q = 'quieter' version, M = modal loudness, L = 'louder version. The shaded area marks weak affective colouring.*

As in the test *apologetic-indignant*, normalisation of the loudness level to modal does not impede affect signalling: all non-modal voice stimuli from Series M yield relatively high affective ratings, with breathy yielding a near-threshold rating.

As suggested by the ICC (r) = 0.78, ICC (R) = 0.98, the raters showed substantial agreement in assigning the ratings to the stimuli.

Relaxed-Stressed

Significant main effects were found for Voice $F(2.35, 35.24) = 85.88$, $p < 0.001$, $\eta_p^2 = 0.85$; Loudness $F(1.08, 16.24) = 17.1$, $p < 0.001$, $\eta_p^2 = 0.53$, and Voice x Loudness interaction $F(4.69, 70.36) = 3.86$, $p = 0.002$, $\eta_p^2 = 0.18$. This suggests that loudness manipulation has different effects on the strength of affective ratings for different voice qualities.

The effect of loudness manipulation on modal and tense voice is similar to the results of the subtests described above: the higher the loudness, the higher the affective ratings for *stressed*; lowering the loudness results in a decrease in the ratings for this affect. The trend is opposite for breathy and whispery voice. For lax-creaky voice the pattern is slightly different: reducing its loudness relative to modal does not change the ratings, while increasing the loudness results in somewhat lower ratings.

Normalisation of the loudness of tense and lax-creaky voice to that of modal voice has no effect on the potency of these stimuli to cue *stressed* and *relaxed* respectively. However, breathy and whispery voice from Series M yield only subthreshold ratings.

The raters' agreement data suggest high degree of coherence in ratings: ICC (r) = 0.77, ICC (R) = 0.98 can be interpreted as substantial agreement.

Sad-Happy

Significant main effects were found for Voice $F(2.1, 31.42) = 22,19$, $p < 0.001$, $\eta_p^2 = 0.6$; Loudness $F(1.19, 17.88) = 27.67$, $p < 0.001$, $\eta_p^2 = 0.65$, and Voice x Loudness interaction $F(4.4, 65.95) = 3.35$, $p = 0.012$, $\eta_p^2 = 0.18$. This again suggests that the effect of loudness manipulation on the strength of affective ratings is voice quality specific.

As in the previous tests, loudness manipulation has similar effect for tense and modal voice, and increasing their loudness entails an increase in affective ratings. It is different for whispery and breathy voice: decreasing the loudness of whispery voice relative to modal (Series Q) is sufficient to yield moderate affective ratings for *sad*, yet this decrease is not enough to yield similar results for breathy voice. Increasing loudness

relative to modal (Series L) lowers the ratings for whispery, breathy and lax-creaky voice, and the effect is less pronounced for lax-creaky.

Normalisation of the loudness to that of modal voice has no effect on the potency of tense and lax-creaky voice to cue *sad* and *happy* respectively. However, breathy and whispery voice from Series M yield low ratings.

The raters' agreement in this test was moderate, as shown by ICC (r) = 0.54, ICC (R) = 0.95.

Scared-Fearless

A significant main effect was found only for Loudness $F(1.07, 16.03) = 12.03$, $p < 0.001$, $\eta_p^2 = 0.45$, the effects of Voice $F(1.36, 20.41) = 2.24$, $p = 0.08$, $\eta_p^2 = 0.13$ was not significant. The two-way interaction $F(3.62, 54.35) = 0.51$, $p = 0.85$, $\eta_p^2 = 0.03$ was not significant either. Although the general pattern of results is similar to that in other tests, the ratings were generally very low (below the threshold). This suggests that none of the stimuli in the set signalled *scared* or *fearless* here, and results will not be further discussed.

As in Loudness experiment 1, the raters' agreement in this subtest was poor: ICC (r) = 0.1, ICC (R) = 0.63.

5.4.5 Summary and discussion

The results of Loudness experiment 2 will be now summarised according to the research questions outlined in the section above.

Q1: How do stimuli differing in voice quality but with loudness normalised yield affective ratings? (Series M)

The ratings of Series M plotted in Figure 5.9 show the effect of loudness normalisation (increase or decrease of the intrinsic non-modal voice quality loudness to that of the modal voice) on voice-to-affect association. In other words, the ratings of the stimuli from Series M in Figure 5.9 allow us to ascertain what the voice quality can achieve when the intrinsic loudness differences are neutralised[26]. It is clear that even with loudness normalisation, non-modal voice qualities are still effective in affect cueing as each is associated with at least one affect with the rating above 1 (which we interpret here as mild to moderate affective colouring). In particular, three voice qualities, tense, lax-creaky and whispery, emerged as the most potent for all affect signalling (no matter what

[26] The intensity level of tense was lowered by 3 dB, the intensity level of lax-creaky was increased by 2.8 dB and the intensity level of whispery voice was increased by about 7 dB (Table 5.4 and Figure 5.8).

level their loudness was set to). They were associated with clusters of affects. Thus, whispery voice from Series M gets associated with *apologetic* and *intimate*, breathy voice – with *apologetic*; lax-creaky voice – with *bored, intimate, relaxed* and *sad*. Note that in cueing *apologetic*, whispery voice gets higher (but not significantly) ratings relative to breathy and lax-creaky voice. In cueing *intimate*, lax-creaky voice is rated higher than breathy voice, but the difference does not reach statistical significance. Tense voice from Series M is associated with *indignant, interested, formal, stressed* and *happy*. The ratings of the modal voice are particularly conspicuous compared to the non-modal voice qualities as in no case did it achieve high ratings for any of the affects. Given that the loudness differences between modal and tense voice are neutralised, it is clear that a tenser voice quality is favoured by the raters as cueing these affects.

Only two affects failed to be rendered by the loudness normalised stimuli from Series M: *fearless* and *scared*. The raters agreement here was poor and the range of ratings for all stimuli in this subtest is broad (see also scatter plots in Appendix 17 and interquartile range in box plots in Appendix 1). This in turn brings the average values to zero. This largely reflects the listeners' uncertainty in voice to affect association in subtest. This was also the case in the cross-language study reported in Chapter 4 and in Loudness experiment 1.

As the original voice quality stimuli with intrinsic loudness differences were not used in Loudness experiment 2, it is impossible to directly compare them and the loudness normalised stimuli and, although informative, the broad indication of voice to affect association is only partially useful. Nonetheless, it is clear that in Loudness experiment 1 and the present experiment the listeners associated the same affect with the same voice quality stimuli, although there is obviously a reduced range of affective ratings where the loudness-normalised stimuli are concerned. It could be concluded that although non-modal voice qualities are still potent in affect signalling, changing their intrinsic loudness level to that of modal voice does influence their potential in communicating affect, as the ratings are somewhat lower.

Q 2: To what extent does loudness variation contribute to affect signalling when stimuli differ in voice quality? How is the affective rating affected by the increase/decrease in loudness level? (Series Q and L)

Significant main effects were found for both voice quality and loudness, as well as for the two-way interaction for all subtests except *fearless-scared* (here the voice quality

and the two-way interaction effects were not statistically significant). This suggests that the effect of loudness manipulation on affective ratings is different for different voice qualities.

As evident from Figure 5.9, the results of loudness manipulation differed for whispery, breathy and lax-creaky on the one hand, and modal and tense voice on the other. In general, increasing loudness of modal and tense voice resulted in an increase in affective rating whereas decreasing the loudness had the opposite effect. For whispery and breathy voice, the pattern is reversed. This trend is evident in all the subtests, and in most cases there is significant difference between the ratings of the stimuli from Series Q and series L. Interestingly, changing the loudness of lax-creaky voice does not appear to have much impact on its affective rating.

Comparison of the ratings of Series Q, M and L in Figure 5.9 shows that the listeners are sensitive to the increase or decrease of loudness in voice quality stimuli.

The effect of loudness decrease (Series Q) is as follows:

(1) for whispery, breathy, lax-creaky voice qualities that are mostly associated with low activation affects, decreasing loudness results in an increase in ratings;

(2) for tense and modal voice qualities that are associated mostly with high activation states, the opposite is true – decrease of the loudness level entails decrease in affective ratings.

Loudness increase (Series L) shows the opposite for the majority of affects: with the loudness increased by 2 dB, tense voice quality receives significantly higher ratings for high activation affects whereas the ratings of lax-creaky and whispery voice decrease. (The only exception here appears to be the *intimate* affect for which the increase in loudness of the lax-creaky voice resulted in virtually no change in affective ratings.)

This is in a way a non-surprising finding as increasing the loudness of tense voice from Series M by increasing its intensity level by 2 dB or decreasing the loudness of lax-creaky or breathy voice from Series M by decreasing their intensity level by 2 dB essentially brings the loudness of these stimuli closer to their original loudness. The results suggest that the closer the stimulus loudness gets to the intrinsic loudness of the original voice quality, the higher the affective ratings, and therefore the more effective affect signalling. An increase or decrease of loudness will only result in the increase of affective ratings for certain voice qualities. On the other hand, when the loudness is not set to extreme values but to that of modal voice, voice quality alone proves to be sufficient for successful affect cueing.

The effect of loudness manipulation is particularly striking for modal voice. Simply increasing its intensity level by 2 dB results in quite a dramatic increase in affective ratings

for *indignant, interested, formal, stressed* and *happy*, even though in no case does the modal_L stimulus achieve ratings higher than those of the tense_L stimulus. The results of the *sad-happy* test are of particular interest. Here modal and tense voice from Series L appear to signal *happy* to the listener-judges. The results suggest that loudness does contribute in signalling the *happy* affect. In the cross-language study reported in Chapter 4, this affect was not found to be cued robustly by modal voice for any of the language groups tested, and tense voice yielded affective ratings for Russian listeners only. Note that the modal_L stimulus was louder than the modal voice stimulus used in the cross-language study, and that the tense_L stimulus had the intensity level lower by about 1 dB compared to the tense voice stimulus used in the cross-language study. It may be the case that this subtle increase/decrease of loudness is important.

In summary, non-modal voice qualities in which the loudness differences have been equalised, are still potent in signalling emotion. The results further suggest that loudness does contribute to affect signalling, but effects are not always symmetrical. For modal and tense, associated with high activation states, increased loudness entails higher ratings, and therefore affective enhancement. Decreased loudness results in lower ratings, though the differences in ratings are less striking. For lax-creaky, and to a lesser extent whispery and breathy, associated with low activation states, loudness decrease brings about affective enhancement: there is a clear tendency towards higher ratings, although less consistent. Loudness increase results in lower ratings. Modal at its original loudness has some association with high activation states (though it never achieves ratings above 1) whereas all other voice qualities at this level do signal some affects. It shows, however, a surprising increase in ratings where the loudness is enhanced (Series L).

5.4.6 Conclusions

The results point towards a conclusion that loudness can contribute positively in affect cueing, particularly when combined with tense-modal voice qualities, increased loudness can enhance signalling of high activation states. It is particularly striking in the ratings of the modal voice, as increasing its loudness by 2 dB results in significant increase in affective ratings for such affects as *indignant, interested, formal, stressed* and *happy*. Here loudness alone appears to achieve affective colouring. For low activation states and associated voice qualities, similar trends show up, though less dramatic and less consistent. Even without intrinsic loudness variation (Series M) non-modal voice quality stimuli prove effective in affect signalling, while modal voice is not.

In view of these results and the results reported in Loudness experiment 1 above it can be concluded that whereas loudness on its own does little to cue affect, but when combined with appropriate voice quality it may be important in the signalling of high activation states. Furthermore, when the loudness approximates the intrinsic loudness of a particular voice quality, it is, perhaps unsurprisingly, at its most effective.

The study supports the hypothesis suggested in Schröder (2004b) that manipulating loudness of a synthesised stimulus while keeping voice quality constant should have a less prominent impact on the stimulus perception than varying the voice quality and keeping absolute loudness unchanged. Loudness appears to work in affect cueing in conjunction with the variations in voice quality.

5.5 Summary

Loudness is defined as the perceived magnitude of the sound. Assumptions of perceived loudness as subjective auditory sensation have to be made based on the results of listening tests using such psychoacoustic procedures as magnitude estimation and magnitude production. Loudness can be expressed in sones (perceived loudness) or phon (loudness level). Objective methods of estimation of perceived loudness include the use of loudness models and loudness meters.

As would be expected, perceived loudness is mainly determined by the sound intensity, but the relationship between sound intensity and loudness is complex. For instance, two sounds being perceived as equally loud may have very different sound intensity (and vice versa) depending on their spectral characteristics and/or bandwidth. The reason for this complex relationship is linked to how sound is processed in the cochlea, i.e. as to whether the acoustic energy is spread over many or only one or a few critical bands (Moore, 2003). Loudness also depends on other properties of the sound, such as its duration and the conditions in which the sound is presented to the listener as well as on the characteristics intrinsic to the listener.

In speech research, loudness has been studied primarily as a perceptual correlate of syllable prominence and stress through acoustic measures related to overall intensity of the speech signal, spectral properties of the signal (spectral slope, spectral balance or spectral emphasis) as well as through the studies of vocal effort. The term loudness has been used with different meaning across studies; and perceived loudness is not infrequently but inaccurately treated as synonymous to intensity. In emotional speech research, loudness,

together with other prosodic features, has been considered an important cue in communicating high activation affects, such as anger and elation.

The chapter focused on role of intrinsic perceived loudness of voice quality in affect cueing. Two experiments were conducted with the synthesised stimuli in which loudness was systematically manipulated. In Loudness experiment 1, stimuli incorporating voice quality features including intrinsic loudness variations were compared to the stimuli where voice quality was kept constant (modal) but in which loudness was systematically modified. In Loudness experiment 2, stimuli incorporating distinct voice quality features all had equal loudness (that of the modal voice). The results of the experiments demonstrated that loudness variation on its own is relatively ineffective whereas variation in voice quality is essential to the expression of affect. In Loudness experiment 1, stimuli incorporating voice quality variations consistently obtained higher ratings than the 'loudness' stimuli. The results of Loudness experiment 2 suggest that even with loudness differences equalised, non-modal voice quality stimuli prove potent in affect cueing. The results further suggest that loudness can contribute positively in affect cueing: when combined with tense-modal voice qualities, increased loudness can enhance signalling of high activation states.

6 Chapter six
Voice source parameters in affect portrayals: an illustrative analysis

6.1 Introduction

As was discussed in Chapter 2 (Sections 2.1.3 and 2.3.3), the introduction of a broader range of parameters in the analysis of the emotionally coloured speech signal is called for in the study of vocal expression of emotions. Johnston and Scherer (2000) and Scherer (2003) argued that the lack of reliable empirical evidence for emotion-specific acoustic patterns could be explained by the exploit of very few relatively simple acoustic parameters (e.g., f_0 range and level or the global statements of intensity). They further argue that measures related to the spectral analysis of the voice source obtained by inverse filtering are likely to provide evidence for better emotion differentiation.

Despite a significant number of studies focusing on the behaviour of the voice source in speech (e.g., Gobl (2003) and references therein), relatively limited data have been presented on how the voice source varies in emotionally coloured speech. The studies reporting such data, e.g., Airas and Alku (2004), Campbell (2004a), Campbell and Mokhtari (2003), Cummings and Clements (1995a), Laukkanen, Vilkman, Alku and Oksanen (1996), Klasmeyer (1997), Waaramaa, Laukkanen, Alku and Väyrynen (2008) have been discussed in Section 2.3.3 of Chapter 2.

This chapter presents a detailed analysis of the voice source variation in a fragment of data elicited for portrayed emotions[27]. It should be emphasised that this is not intended to be a description of genuine, naturally occurring emotion in speech: what findings are reported here pertain only to the limited kind of portrayals analysed here. Furthermore, as the study involves single utterances produced by a single speaker; these results can only be illustrative of how these portrayals may vary at the level of the voice source, and are clearly not representative of the range of possibilities in portrayal even by one speaker.

The main purpose of this analysis was to illustrate how a rich source analysis (including perhaps redundant parameters) might allow detailed acoustic characterisation of affects. It should be noted that rather than focusing the measurements on short steady state vowel segments (as in most studies concerned with the voice source variation in affective speech, e.g., (Airas & Alku, 2004; Toivanen et al., 2006; Waaramaa et al., 2008), a sentence of several syllables was analysed in much detail. Two issues were of particular

[27] This chapter is an extended version of the following paper: Yanushevskaya, I., Tooher, M., Gobl, C., & Ní Chasaide, A. (2007). Time- and amplitude-based voice source correlates of emotional portrayals. In A. Paiva, R. Prada & R. W. Picard (Eds.), *Affective Computing and Intelligent Interaction: Proceedings of the ACII 2007* (Vol. 4738, pp. 159-170). Lisbon, Portugal: Springer-Verlag.

interest. One was to look at whether (even for a single speaker and limited data such as these) emotion-specific settings of glottal parameters can be established with this kind of analysis. Furthermore, it was hoped that the analysis, by looking in great detail at even a single utterance might shed light on the dynamic aspects of source variation. This is often claimed to be crucial in affective speech, but there exists almost no data to support this at the level of the voice source.

As discussed in Chapter 2, the general movement in the field now is towards the use of naturalistic affective speech data or induced emotion expressions, while acted expressions are generally viewed as undesirable and are used increasingly less frequently (but see, for example, Bänziger and Scherer (2007)). Of course, the study of true emotion will require genuine emotion expression (as in naturalistic corpora or at least obtained with emotion induction procedures). For the kind of detailed voice source analysis used in this study, the available expressive data is largely unsuitable. The reasons are described in Chapter 2 and concern primarily (i) recording conditions; (ii) lack of control over the segmental and prosodic context.

While fully acknowledging a somewhat problematic ecological validity of simulated data, we opted for the use of simulated emotions for the purposes of our analyses for the following reasons.

The data analysis involved inverse filtering of the utterances to obtain the differentiated glottal flow and subsequent LF model matching to obtain salient source parameters. For this type of analysis to yield meaningful voice source data, the recording conditions have to be very stringent. Controlling for segmental and prosodic content is another important consideration: segmental and prosodic characteristics have very large effects on the source (Ní Chasaide & Gobl, 2004a). As outlined in Chapters 1 and 2, it is difficult to separate the effects of emotion and affect from these other effects.

These considerations led to the use of portrayed rather than naturally occurring expressions of emotion. These portrayals do have some relation to real life use of voice. For example, they illustrate the kinds of voice source variations that might be used for narration, e.g., when reading a bedtime story to a child. In everyday life, people often choose to purposefully vary their voice to communicate (often for manipulative purposes) a certain affective message. Whereas sometimes we reveal ourselves intentionally or unintentionally through our voices, quite a lot of the time we are adopting voice qualities to mislead the listener. For example, we feign interest while often bored in a conversation or commiserate with someone's woes with a sad voice that belies our secret indifference. In

other words, portrayals are very much part of the real world of human expressive communication.

Actor speech was purposely not chosen for our recording. It was felt that the stylistic devices used by stage actors tend to be even further removed from the real life analogies referred to above. While actors might tend to produce exaggerated portrayals, non-actors (we hoped) would be likely to modify their voice to reflect such everyday use of the voice. No claims are made here that the recorded portrayals used in this study are representative of genuine emotional vocalisations. Rather, they exemplify voice source variations that might be used for narration.

6.2 Material and method

6.2.1 Speech sample

The recorded data consisted of repetitions of an all-voiced utterance 'We were aWAY a YEAR ago' produced by a male speaker of Hiberno-English, a non-actor. The recording was made in a semi-anechoic chamber with the Pearl CC30 condenser microphone. The distance from the microphone was kept constant at 30 cm. The signal was recorded directly to computer using the MultiSpeech software package, at the sampling frequency 44.1 kHz.

The speaker rendered the following affective states: *sad, happy, bored, surprised, angry, afraid, disgusted* as well as *neutral.* This set includes affective states most extensively studied in the field of the vocal expression of emotion, and was chosen for this case study in part to complement the existing empirical data with a detailed analysis of the voice source characteristics of a speech sample in which these affective states were portrayed.

In the course of the recording, the speaker was advised to keep the peaks of prominence (accents) on the same syllables (WAY and YEAR) in all repetitions. For each portrayed emotion, at least four repetitions were recorded. On the basis of the subsequent auditory analysis, one repetition was selected for the instrumental analysis. Given the type of analysis (inverse filtering), the choice of the utterance was influenced by the overall quality of the signal and the authenticity of the emotional portrayal.

To confirm that the recorded portrayals indeed represented the targeted affective states, a listening test was conducted.

6.2.2 Listening test

The stimuli for the listening test included the 7 portrayals of affective states (*bored, sad, happy, surprised, angry, afraid* and *disgusted*) and 3 repetitions of the *neutral* utterance, 10 in total. The sound files with the stimuli were presented to the listeners on a computer screen in randomised order, 5 randomised lists were generated. 16 volunteers participated in the test. The participants were asked to listen to each of the sound files as many times as they wanted, to determine which affect is expressed in each sound file, and to mark their choice by clicking the radio button next to the affective label listed next to the sound file. The affective labels included *happy, sad, bored, angry, surprised, disgusted, afraid* and *no emotion*. A blank box was also added to the list of labels so that the listeners could provide their own affective label should none of the available labels prove adequate. The results of the listening tests are presented in Table 6.1.

Table 6.1. *The results of the perception test (%). Emotions recognised as targeted in 70% or more cases are shown in bold type.*

Perceived as / Target	Sad	Happy	**Bored**	**Surprised**	Disgusted	**Angry**	Afraid	**Neutral**	Other
Sad	58	-	8	-	1	-	6	21	6
Happy	-	46	-	5	5	12	2	28	2
Bored	18	-	**71**	-	-	-	1	9	1
Surprised	-	6	-	**93**	-	-	1	-	-
Disgusted	1	1	12	6	24	-	-	51	5
Angry	-	1	1	1	2	**91**	-	2	2
Afraid	56	1	1	-	2	1	28	6	5
Neutral	8	3	6	-	6	2	2	**71**	2

The analysis of the listening test showed that the targeted affects were recognised at a level of 70 per cent or higher in 4 out of 8 portrayals. The most readily recognised were *surprised* (93 per cent) and *angry* (91 per cent), followed by *bored* and *neutral* (both with 71 per cent 'correct' recognition). *Bored* was also perceived as *sad*, but only in 18 per cent of cases. *Sad* received somewhat lower recognition rates: the listeners recognise it as such in 58 per cent of cases. *Sad* was identified as *neutral* in 21 percent of cases. *Happy* also received relatively low recognition rating, 46 per cent, and it could be confused principally with *neutral* (28 per cent) or *angry* (12 per cent). *Afraid*, recognised as such in only 28 per cent of cases, was most readily perceived as *sad* (56 per cent). *Disgusted* got the lowest recognition rates compared to the other emotions, only 24 per cent listeners identified it as such, and it was identified primarily as *neutral* (51 per cent). As disgust is more often than

not expressed in affect bursts rather than in longer utterances and with facial expression acting as a strong cue, its low recognition rates were somewhat expected.

The utterances that were recognised by the listeners as conveying the targeted emotions in more than 70 per cent of all cases were selected for the further analysis involving inverse filtering. Despite the relatively lower recognition rates, both *sad* and *happy* were included in the voice source parameter analysis, in order to represent a broader range of affective states.

6.2.3 Data analysis

The analysis of the recorded material was performed using inverse filtering and voice source model matching techniques. As discussed earlier (Chapter 2), in inverse filtering, the speech signal is filtered with the inverse of the vocal tract filter characteristics. This cancels the filtering effect of the vocal tract transfer function, and the estimation of the differentiated glottal source signal is thus obtained. The obtained glottal waveform is then matched to one of the existing models; in this case, the LF model (Fant, 1995; Fant et al., 1985) which allows one to obtain important voice source parameters. Details of these techniques are given in Chapter 2. The tool used for the present analysis is described in detail in Monahan (1996), Ní Chasaide, Gobl and Monahan (1992) and Gobl and Ní Chasaide (1999a), see also Figure 6.1.

The selected six utterances (*bored, sad, happy, surprised, angry* and *neutral*) were inverse filtered. Prior to inverse filtering, each sound file was resampled at a sampling frequency of 10 kHz and high-pass filtered using a phase linear high-pass filter with a cut-off frequency of 40 Hz, to ensure the correct zero pressure line. It should be noted that for the listening tests, the re-sampled and high-pass filtered utterances were used.

Each utterance was initially inverse filtered using automatic inverse filtering software based on closed phase covariance LPC, to obtain a first estimate of the differentiated glottal flow. Subsequent manual, interactive fine-tuning of the inverse filter was performed, pulse by pulse, for all utterances. The details of the system are described in Monahan (1996), Ní Chasaide, Gobl and Monahan (1992) and Gobl and Ní Chasaide (1999a). The number of pulses in the utterances varied depending on affective state, e.g., there were 81 pulses in *bored* and *neutral*, 86 in *sad*, 96 in *surprised*, 105 in *happy* and 129 in *angry*. Interactive software (Gobl & Ní Chasaide, 1999; Monahan, 1996; Ní Chasaide et al., 1992) was also used to manually fit the LF model to the inverse filtered waveform to obtain measures of voice source parameters.

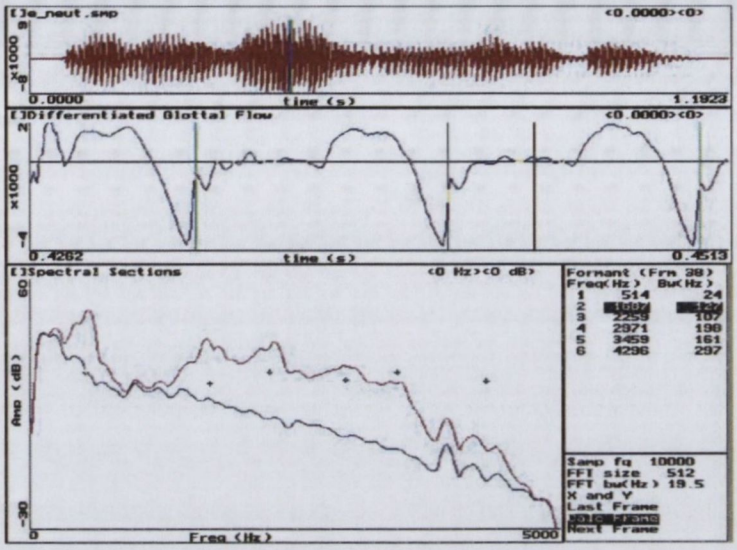


Figure 6.1. *Illustration of the inverse filtering software tool.*

The following voice source parameters were included in the analysis: F0, EE, RK, RG, RA, FA, OQ and RD. These parameters were described in detail in Chapter 2 (see Table 2.1), and are briefly glossed here again in Table 6.2. More detailed descriptions of these parameters and their spectral correlates are also given in Gobl (2003), Gobl and Ní Chasaide (2003b) and Ní Chasaide and Gobl (1997).

Table 6.2. *Voice source parameters selected for the analysis.*

Parameter	Brief description
F0	fundamental frequency, 1/T0 where T0 is the duration of the glottal pulse
EE	a measure of the strength of the main glottal excitation
RK	a measure of the skew of the glottal pulse. Higher RK - more symmetrical glottal pulse
RG	the glottal frequency FG normalised to F0, where FG is the characteristic frequency of the glottal pulse during the open phase
RA	Ta/T0, where Ta is a measure of the effective duration of the return phase after the main excitation, prior to glottal closure. Related to spectral tilt
FA	a parameter related to RA, also a measure of spectral tilt. High FA value - source spectrum with relatively strong higher harmonics
OQ	the duration of the glottal open phase in relation to the duration of the whole glottal period (can be derived from RG and RK)
RD	a global waveshape parameter derived from F0, EE and UP, where UP is the peak amplitude of the glottal flow pulse

6.2.4 Statistical analysis and data processing

For each of the analysed utterances, mean and standard deviation of each parameter (F0, EE, RK, RG, RA, FA, OQ and RD) were calculated (Figure 6.2). To explore parameter variability as a function of affective state, a one-way ANOVA with subsequent Tukey's HSD test was conducted, for each parameter separately. The results of the ANOVA are given in Table 6.3; the results of the post-hoc Tukey's HSD test are given in Table 6.4. Parameter values were found significantly different statistically at the $p < 0.05$ level across all the affective states; see Table 6.3. High eta squared (η^2) values are suggestive of significant effect of affective state: over 40 per cent of the variance in mean parameter values is explained by this variable.

The mean values for parameters in Figure 6.2 only give a very global picture of variation found for the different emotions. To better capture the parameter dynamics across the duration of the utterance, the rate of change in each parameter was obtained by calculating the first order difference from the smoothed parameter values. Figure 6.3 shows the mean and standard error of the absolute rate of change for the analysed glottal parameters.

In order to facilitate visual inspection of the dynamics of each affect-coloured utterance (Figures 6.4 and 6.5), the time axis of each utterance was normalised to that of *neutral*. A number of anchor points in the *neutral* utterance were chosen for the time normalisation. These initially included anchor points at syllable boundaries. For vowels from the accented syllables, additional anchor points were placed at their midpoints as well as at the approximate boundary of [w] and [e] in the stressed syllable WAY. The [g] segment was excluded from the analysis due to the differences in its realisation in terms of voicing across utterances. In Figures 6.4 and 6.5 the anchor points are shown as vertical lines. For each part of the utterance between anchor points, the time axis was scaled to be of the same duration as that of the corresponding *neutral* one. As the utterances also had a different number of pulses, linear interpolation was performed in order to plot all affective utterances to the same time axis points as that of the *neutral* utterance.

To reduce noise caused by small pulse-to-pulse parameter variation, a moving average of parameters was calculated. The frame spanned 3 pulses with a 1 pulse frame-shift. This served to smooth the trajectories, while preserving the overall parameter dynamics. The resulting illustrations of parameter dynamics across the various affectively-coloured are represented in Figures 6.4 and 6.5.

Table 6.3. *Results of the one way ANOVA for parameter values.*

Parameter	df	F	Sig.	Eta^2
F0	5, 665	136.94	<0.001	0.51
EE	5, 665	126.39	<0.001	0.49
OQ	5, 665	335.49	<0.001	0.72
RD	5, 665	263.74	<0.001	0.66
RA	5, 665	113.43	<0.001	0.46
RG	5, 665	162.20	<0.001	0.55
RK	5, 665	96.51	<0.001	0.42
FA	5, 665	172.36	<0.001	0.56

Table 6.4. *Results of Tukey's HSD test comparing parameters across emotions. .Asterisks show significant difference at $p < 0.05$. Values representing overlap between activation groups are shown in bold type.*

F0	neu	bor	sad	hap	surp	**RA**	neu	bor	sad	hap	surp
bor	1.00					bor	0.52				
sad	0.57	0.50				sad	*	*			
hap	*	*	*			hap	**0.09**	*	*		
surp	*	*	*	*		surp	*	*	*	*	
ang	*	*	*	*	0.08	ang	*	**0.38**	*	*	*
EE	neu	bor	sad	hap	surp	**FA**	neu	bor	sad	hap	surp
bor	0.26					bor	*				
sad	*	*				sad	*	*			
hap	*	*	*			hap	**0.79**	*	*		
surp	**1.00**	**0.25**	*	*		surp	**0.97**	*	*	1.00	
ang	*	*	*	0.99	*	ang	*	*	*	*	*
OQ	neu	bor	sad	hap	surp	**RG**	neu	bor	sad	hap	surp
bor	0.34					bor	0.99				
sad	*	*				sad	*	*			
hap	*	*	*			hap	**0.92**	**0.39**	*		
surp	*	*	*	*		surp	*	*	*	*	
ang	*	*	*	*	*	ang	*	*	*	*	*
RD	neu	bor	sad	hap	surp	**RK**	neu	bor	sad	hap	surp
bor	0.85					bor	*				
sad	*	*				sad	*	*			
hap	*	*	*			hap	*	*	*		
surp	*	*	*	*		sur	*	*	*	0.40	
ang	*	*	*	*	*	ang	*	*	*	1.00	0.12

6.3 Results and discussion

The ANOVA results are shown in Table 6.3. Table 6.4 details the results of Tukey's HSD test. Figure 6.2 shows the mean and standard deviation for each of the parameters across emotions. Figure 6.3 shows the mean and standard error values of the absolute first order difference (rate of change) of glottal parameters. Figures 6.4 and 6.5 illustrate the dynamics of parameters across emotions.

According to the level of activation, the affective states portrayed in the utterances analysed here fall into two distinct groups: low activation *neutral*, *bored* and *sad*, and high activation *happy*, *surprised* and *angry*. The discussion of results will be focused on each of these activation groups.

The post-hoc pairwise comparisons (Table 6.4) showed that for this limited dataset, parameters F0, OQ, RD and RK differentiate well between high and low activation groups. RK has been found to best differentiate between activation groups, as well as among affective states within the low activation group. The results observed for F0 were somewhat expected, as fundamental frequency is an established correlate of affect (see discussion in Chapter 2). OQ and RD, a global waveshape parameter, were found to differentiate among all affective states portrayed in the utterances analysed here, with the exception of *neutral* and *bored*. A comparison of the RD data obtained in the present experiment with the NAQ (essentially the same global waveshape parameter as RD) values reported in Airas and Alku (2004) showed similar patterns for both global measures. In Airas and Alku (2004), NAQ values for all affective states except *angry* were higher than *neutral*. Similar patterns were observed here for RD, except *happy* was also lower than *neutral*.

EE, RA, FA and RG were not significantly different for affective states from the two different activation groups. For example, EE of *surprised* is not significantly different from that of *bored* and *neutral*, and RA for *angry* and *bored* yielded similar values. This overlap between the activation groups in EE, RA, FA and RG suggests that one parameter is unlikely to provide sufficient information about the voice source in certain affective states and that these parameters in particular should be considered in conjunction with other voice source parameters. The analysis of parameter values suggests that within-group affective state differentiation is parameter specific, with some parameters distinguishing among affective states more clearly than others. This differentiation has been found here to be activation-group specific. Within the high activation group, affective states are differentiated in terms of OQ, RD, RA and RG, whereas within the low activation group the only differentiator appears to be RK. All voice source parameters except RK showed

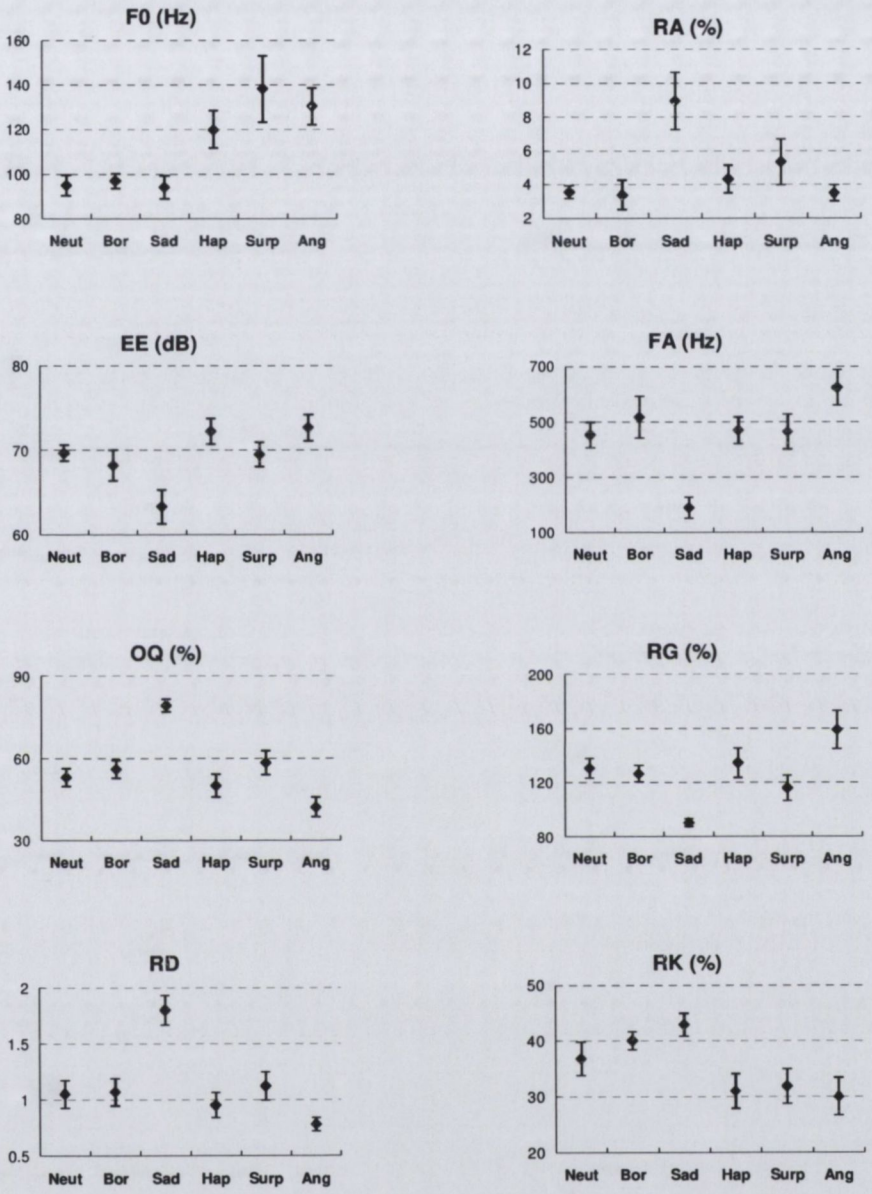


Figure 6.2. *Mean and standard deviation of glottal parameters across emotions.*

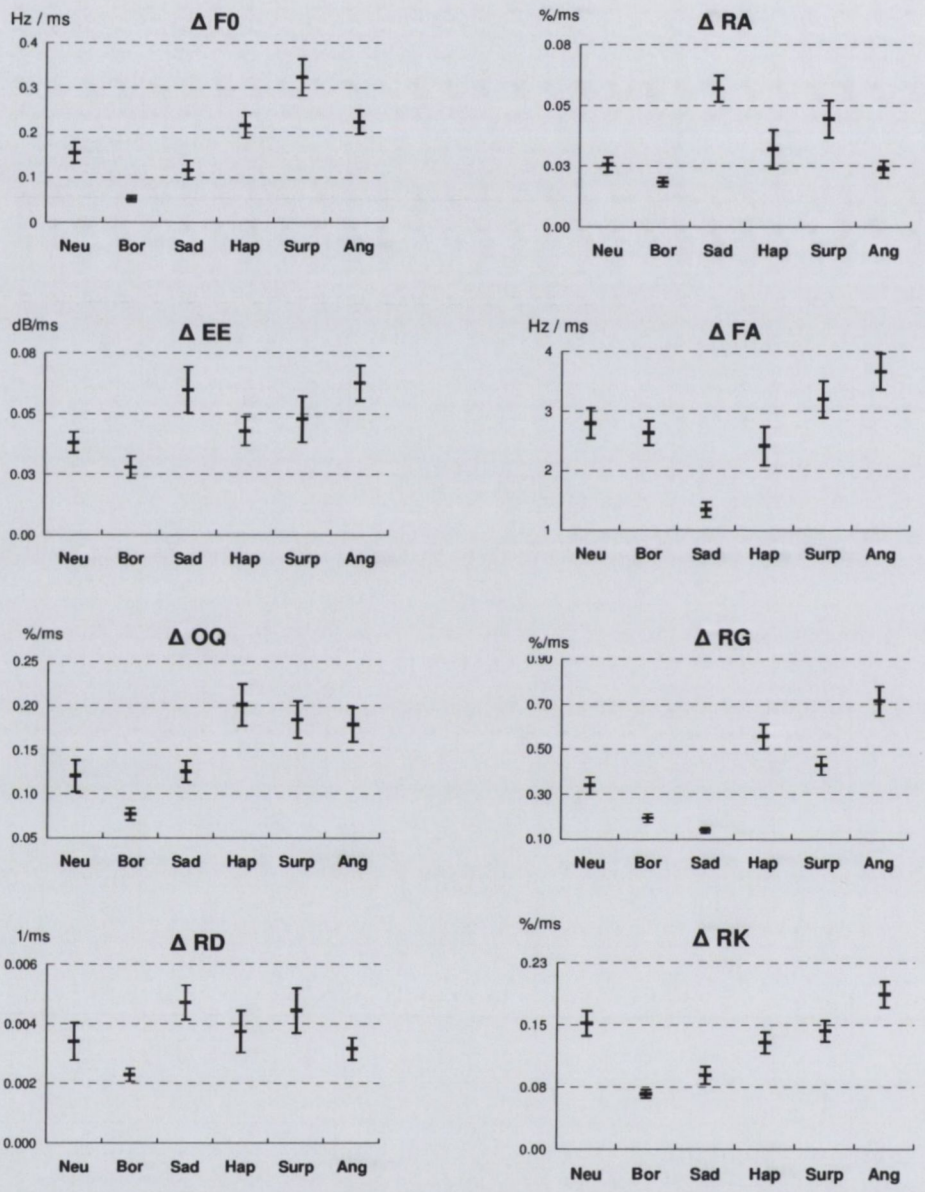


Figure 6.3. *Mean and standard error of the rate of change of parameters across emotions.*

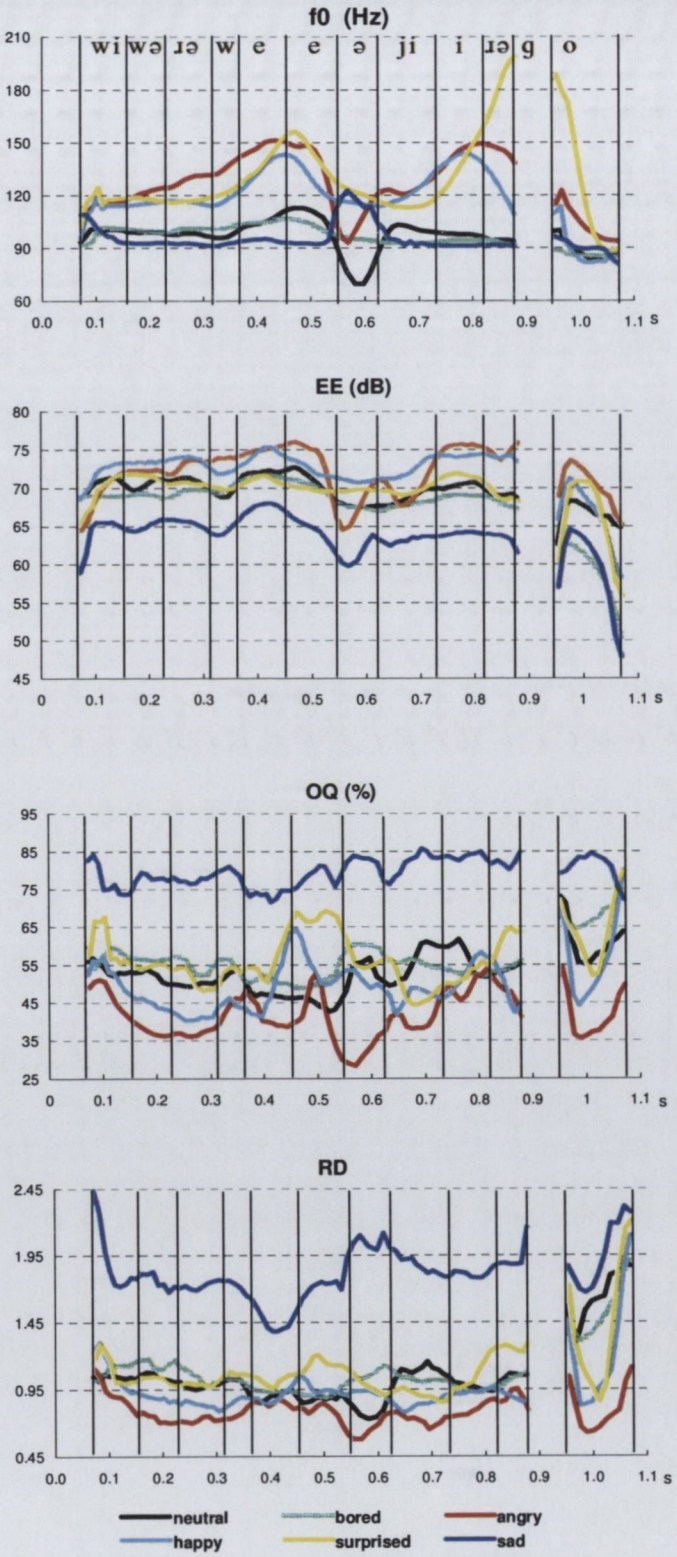


Figure 6.4. *Dynamics of glottal parameters.*

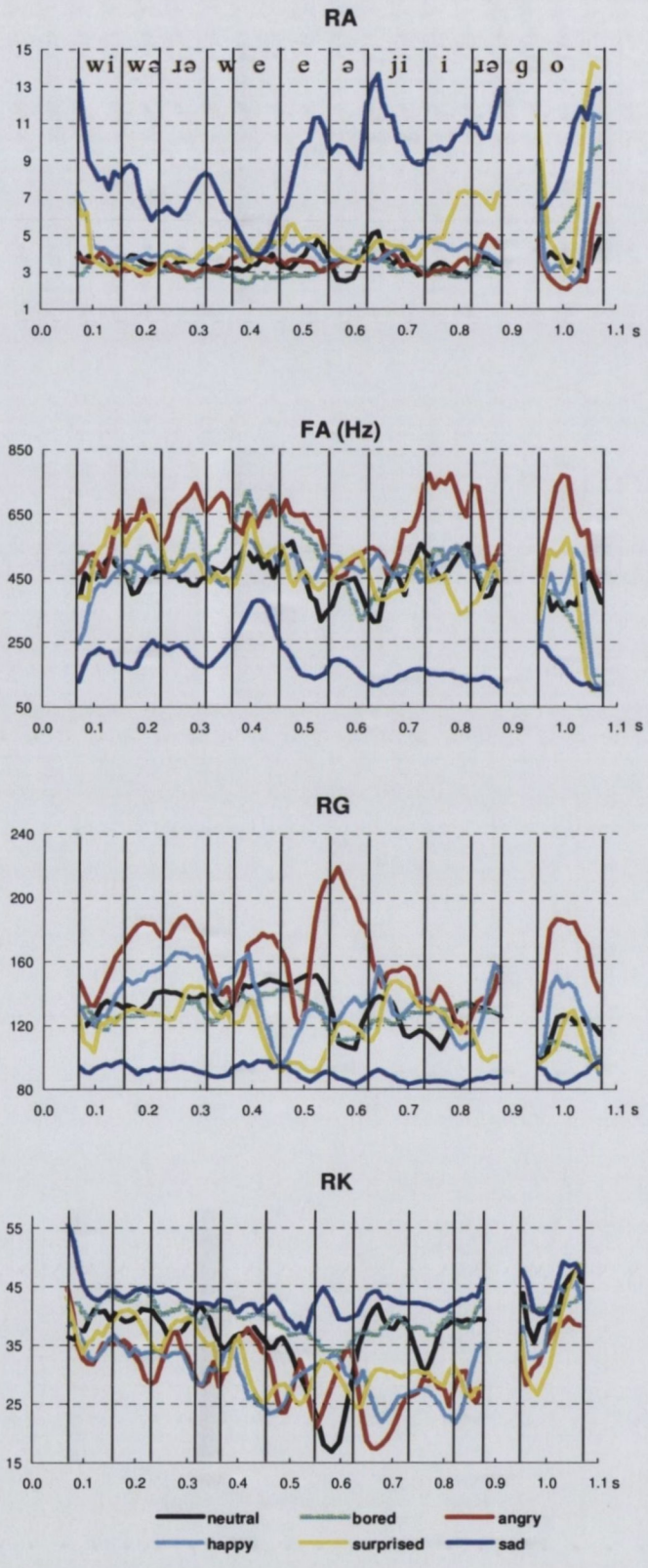


Figure 6.5. *Dynamics of glottal parameters (continued).*

no statistically significant difference between *bored* and *neutral*, which in itself is noteworthy.

It is obvious both from the mean values in Figure 6.2 and from the smoothed parameter trajectories in Figures 6.4 and 6.5, that the voice source parameter values of *angry* and *sad* almost always appear as extremes (highest and lowest), each representative of the most extreme affective state within their relevant activation group. It is interesting to note that these affective states have been associated with very different voice qualities, which is supported by the combination of parameter settings found here.

The analysis of the present data sample revealed that when parameters show no affect differentiation in terms of mean values, it is the parameter dynamics that distinguish affective states. For example, *bored* and *neutral* yield similar mean values for the majority of parameters measured here (except RK), but show very different rates of change, parameters for *neutral* being of more dynamic nature. Within the high activation group, for example, *angry* has not been found to be differentiated from *happy* and *surprised* in terms of mean RK, but it shows markedly higher rate of change for this parameter.

Further inspection of the data draws attention to the great degree of utterance-internal dynamic variation of all parameters and in all renditions. This in itself is of course not a new finding, and this kind of source variation has been discussed earlier (e.g., Ní Chasaide & Gobl, 2004a; Ní Chasaide & Gobl, 2004b). For example, the inspection of the parameter trajectories over the utterance shows that the most stable source values tend to be associated with the stressed vowels, while the final (unstressed) syllable of the utterance tends to show extensive source fluctuations. This suggests that the critical affect-related differences are likely to be linked to prosodically important parts of the utterance.

The global trends in parameter dynamic variation are represented here as mean delta values. It is clear, however, that the picture is a lot more complicated. Figure 6.6 represents delta values compared for stressed and unstressed syllables of the utterance for the parameters analysed. It is obvious from Figure 6.6 that the rate of change varies as a function of the syllable prominence. (Note also that a voice source difference between stressed and unstressed syllables is not a new finding and has nothing to do with affect per se.) In general, parameters show higher stability in the stressed syllables compared to the unstressed ones. On the other hand, the syllable prominence appears to have different impact on dynamic variation of parameters in the different affective states portrayed in the short sample here. Thus, for example, the rate of change of F0 has been found to be markedly different for the two types of syllables in *surprised*, being higher in the

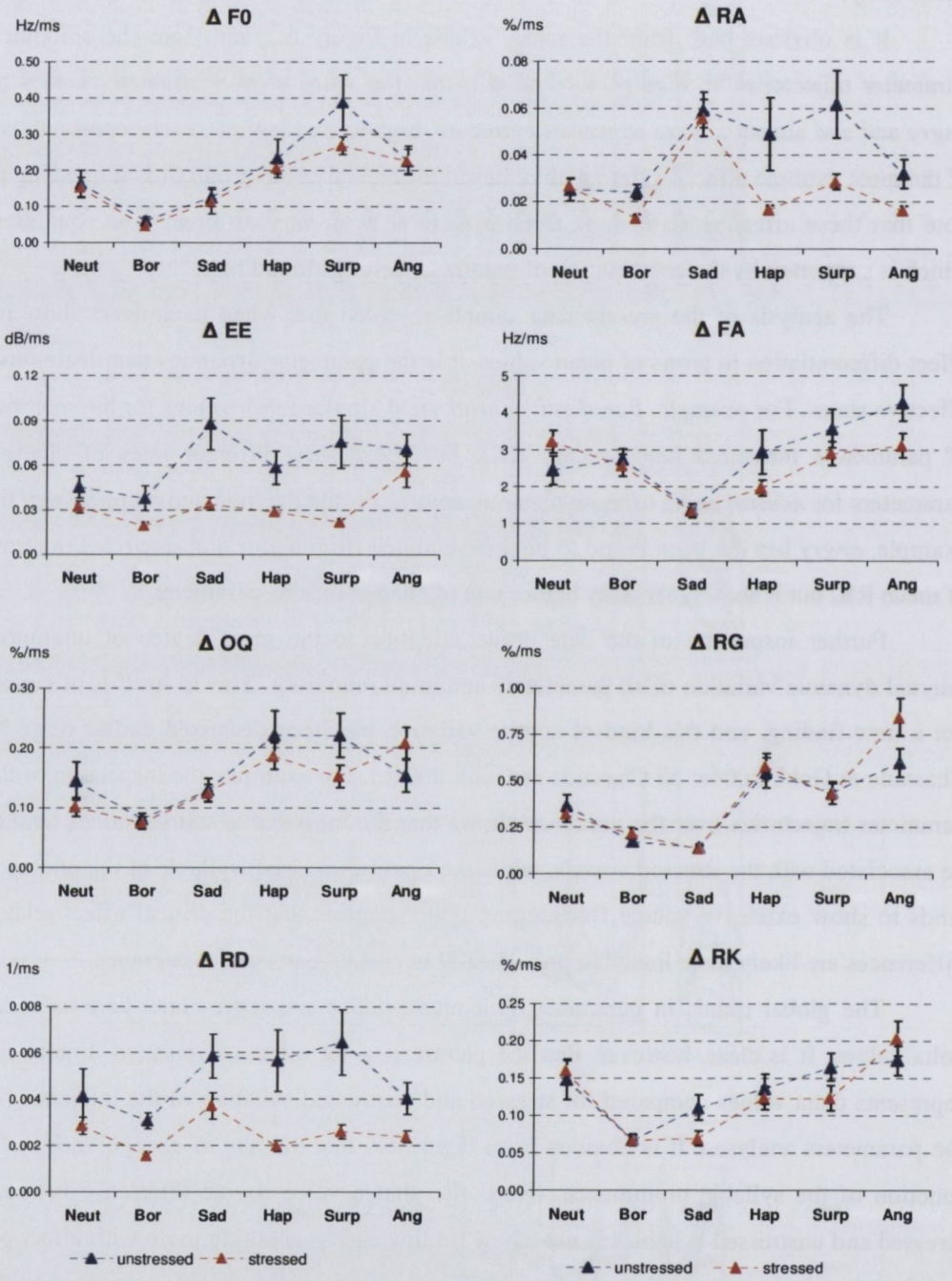


Figure 6.6. *Mean and standard error of the rate of change of parameter values across emotions: stressed and unstressed syllables compared.*

unstressed syllables, whereas in other affective portrayals analysed here this difference is only minor. In *neutral, bored,* and *sad,* for a number of parameters, such as F0, OQ, RA, RG, RK and RK, parameter dynamics do not seem to be affected by syllable prominence. For the affects from high activation group, this difference is more pronounced.

Close examination of the dynamic aspects (rate of change values) of the global waveshape parameter RD shows that, as already mentioned, the dynamic variation is much greater on the whole in the unstressed syllables. Across the different affective portrayals analysed here, there are major differences in (a) the mean delta values for RD: for example, *bored* has consistently lower dynamic variation than the others, including *neutral*, and (b) the difference between unstressed and stressed syllables in terms of the stability of dynamic variation (shown here as the standard error of the mean delta values): for example, for *happy* and *surprised* the difference is much more striking than for *bored, angry* and *neutral* portrayals.

The results suggest that the affect related dynamics are crucial to understand how the voice portrays emotion. These affect related dynamics appears to sit on top of the considerable prosodically determined source dynamics, which have little to do with affect expression per se. To understand affect-related voice source dynamics we will need in parallel to grow our understanding of the underlying prosodic dynamics as the affect related shifts in the former are added to the latter.

A summary of the combinations of parameter settings for each of the portrayed affects in the sample analysed here is shown in Figure 6.7. Parameter levels were calculated as a percentage difference relative to *neutral* based on mean values, and the ranges represented here are somewhat arbitrary: -2 = [< -25%] (very low), -1 = [-25%, -5%] (lower than neutral), 0 = [-5%, 5%] (within the neutral range), 1 = [5%, 25%] (higher than neutral), 2 = [> 25%] (very high). Colour-coded circles show parameters that have high dynamic variation (as shown by mean delta values in Figure 6.3).

As evident from Figure 6.7, the combination of parameter settings is different for each of the affective portrayals in this short sample. For example, *neutral* and *bored* are differentiated in terms of the level of RK and RA, and *bored* and *sad* are distinguished by all parameter levels except F0. It is particularly conspicuous that no parameter in *bored* shows high dynamic variation, which further contrasts the settings for this emotion to those of *sad*. These results suggest that it is highly unlikely that the voice source difference among affective states could be captured by a single measure or a combination of static parameters. A combination of dynamically varying parameters will need to be considered.

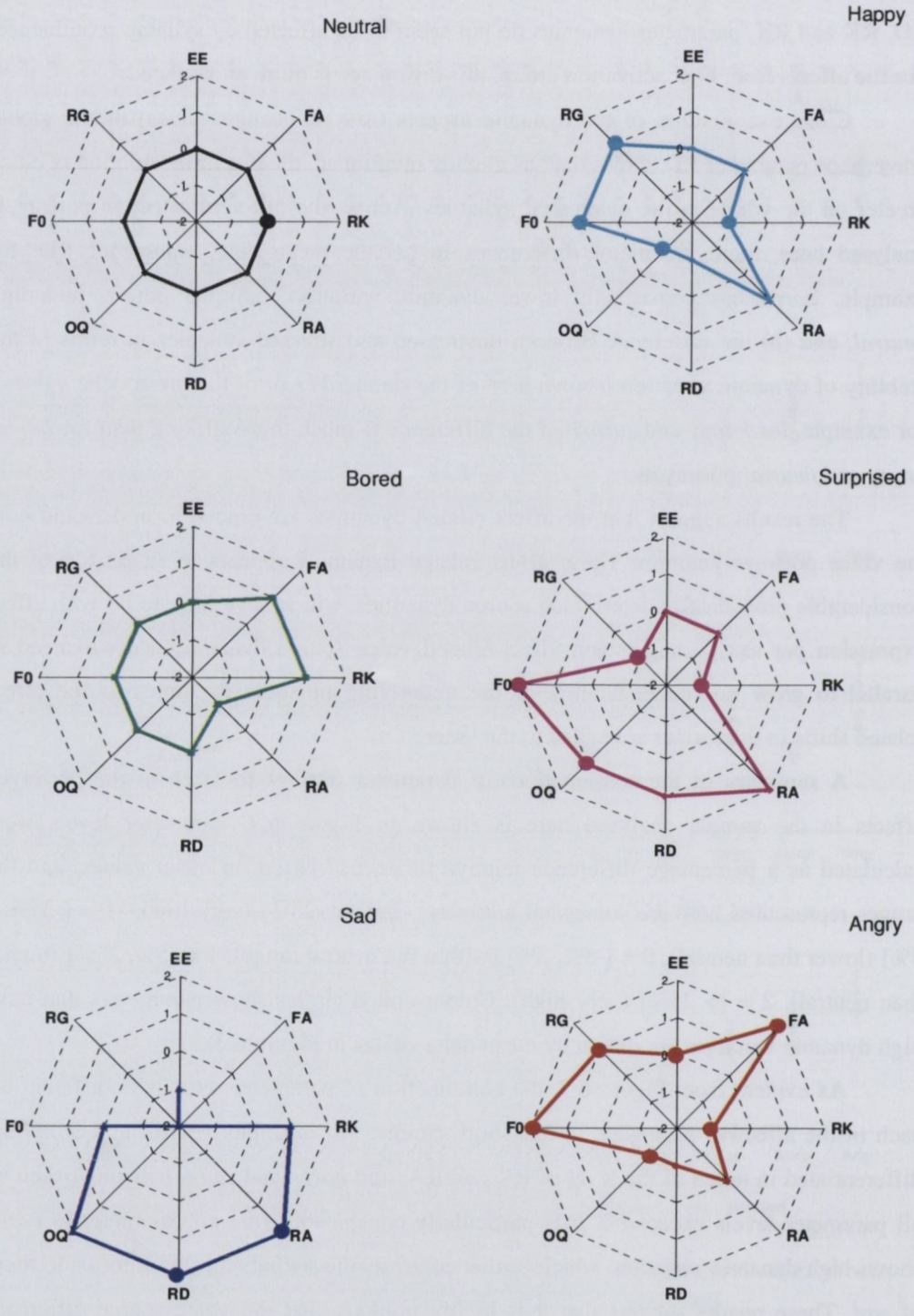


Figure 6.7. *Sample global profiles of parameter settings: summary. -2 = [< -25%] (very low), -1 = [-25%, -5%] (lower than neutral), 0 = [-5%, 5%] (within the neutral range), 1 = [5%, 25%] (higher than neutral), 2 = [> 25%] (very high). Colour-coded circles show parameters demonstrating high dynamic variation (high delta values).*

6.4 Conclusions

Voice source parameters were analysed in a limited dataset that consisted of utterances portraying a number of different affective states, in the production of one speaker. This study was intended as an illustration of how a rich voice source analysis (including a number of salient voice source parameters) might allow acoustic characterisation of the different portrayed affects. The parameters chosen for analysis were those that give an overall picture of the glottal pulse.

This study is not and was not intended to be an analysis of true affect. It was highly unrepresentative in lots of ways: it involved portrayed affect rather than truly occurring affect; it involved a single speaker; the analysis was carried out on only one utterance for each of the portrayed affects. To this extent, the results are not presented to make any kind of definitive claims about how affects differ. Rather, it was intended as an illustration of how a detailed parameter-rich analysis can proceed, and intended to explore in a preliminary way the kinds of dimensions that may be important in affect differentiation.

Although the conclusions can only be made for this very limited speech sample, this study yielded some insights. One was that even broad averaged representations (as in Figure 6.7) may allow differentiation of these portrayed affects. For example, parameters that differentiate well between activation groups are F0, RK, RD and OQ. It is clear, however, that what is presented here is only a preliminary first step. Much more data would be required to come up with more robust characterisations.

The study serves to highlight that while these global characterisations can give a useful overview of vocal settings, they provide in themselves only a limited insight into how the voice varies with affect in speech. As often mentioned (Fónagy, 2001), the comprehensive description will need to grasp the utterance-internal source dynamics. The brief analysis in Chapter 6 draws attention to some of the dynamic aspects of source variation in these utterances. One thing to emerge clearly is that the differences between certain affective renditions lie much less in their averaged vocal parameter settings than in the dynamic variation of the parameter. So, for example, average RD values for *neutral* and *bored* are almost identical, but the dynamic variation in *bored* is dramatically less.

The analysis drew attention to a very considerable within-utterance dynamic variation. Much of this variation is not related to affect per se but rather to the underlying prosodic patterning: the results suggest that important aspects of dynamic variation seem to be linked to certain points of the utterance, i.e. most likely the prosodically important chunks (such as

stressed/unstressed syllables). The across-portrayed affect variation appears to sit on top of this prosodic patterning of the voice source. To understand the affect-related voice source variation, it appears clear that we need to build in parallel the understanding of the underlying prosody of the voice.

The results of this case study suggest that parameter-rich voice source characterisation can be useful to the description of affect-specific voice quality patterns. This kind of description is likely to be of relevance for the generation of expressive synthesis. In a recent work reported in Tooher et al. (2008) the voice source parameter settings for the speech sample described here have been used in a resynthesis of emotionally coloured utterances using classification and regression trees. The approach to voice source parameter dynamics modelling reported in Tooher et al. (2008) involved building regression trees for each voice source parameter using the parameter settings of the *neutral* utterance in order to transform the voice source parameter trajectories from *neutral* to those of *bored, sad, happy, surprised* or *angry.* In the resynthesis of utterances, the input of each parameter regression tree consisted of the target emotion, the source parameter values of *neutral*, and the information on syllable prominence (stressed/unstressed). The output of the regression trees were the new emotion-specific source parameter values that were subsequently given as input to the KLSYN88a synthesiser, along with the vocal tract filter data of the *neutral* utterance. The results of both formal and informal listening tests showed reasonable success of this approach in that the intended affective colouring was achieved in the resynthesis. Further work is necessary, however, mostly related to the modelling of dynamic variation of the voice source parameters. For example, the accuracy of transformed source parameters trajectories could be further improved by including more detail to the coding of syllable prominence in building and use of the regression trees. The information on syllable prominence (stressed/unstressed) could be enhanced by considering its role in the prosodic organisation of the intonation phrase, such as whether the syllable represents pre-nuclear or nuclear accents, or whether it is pre-nuclear or post-nuclear unaccented syllable. This again highlights the need to develop in parallel an understanding of the underlying prosody of the voice, if we want to get the full grasp of the affect-related voice source variation.

7 Chapter seven
Conclusions

7.1 Introduction

The main motivation behind this study was to broaden the knowledge related to the way voice quality is used in affect communication, and in particular, with regard to what is universal and what is language-/culture-specific in the way the voice is used to signal affect.

This thesis presents results of a series of synthesis-based studies exploring perceptual aspects of voice quality in the communication of affect as well as an analytical case study of the voice source variation in an affectively coloured speech sample. The following sections outline the main findings of the study as well as some suggestions for future work.

7.2 Main findings and suggestions for future work

7.2.1 Cross-language study

The cross-language study was the main study of this thesis. It aimed, in a series of synthesis-based perception experiments and for a limited number of different affective states, to clarify to what extent voice quality can impart different affective colouring for subjects from four different language backgrounds (Hiberno-English, Russian, Spanish and Japanese). The study also investigated to what extent this affective colouring can be enhanced by including f_0 contours with different level and dynamics in some stimuli, particularly for the signalling of strong emotions.

The results of the study showed broad cross-language similarities in the way voice maps to affect, as well as some major cross-language differences.

For all languages tested, all non-modal 'VQ only' stimuli get associated with a number of affects. In most cases, there is no one-to-one voice-to-affect mapping. Similarly, the same affect can be cued by more than one voice quality. Overall, for the range of stimuli included in this experiment, those which incorporate distinct voice qualities ('VQ only' or 'VQ + f_0' series) emerged as being the most effective in signalling affect. The 'f_0 only' stimuli appear to be relatively less effective. This holds for all the languages tested, and there is considerable cross-language agreement in the way these stimuli signal quite a number of affects (*sad, scared, indignant, apologetic, interested, bored, fearless*).

As for the cross-language differences, these were particularly prominent for the subtests *intimate-formal* and *relaxed-stressed*. The *intimate* affect was associated for Spanish and Japanese subjects with tense/modal voice combined with f_0 'indignation', a stimulus which was associated rather with *formal* for the Hiberno-English and Russian subjects. On the other hand, *intimate* was cued by lax-creaky voice, lax-creaky + f_0 'boredom', whispery voice and whispery + f_0 'fear' for Hiberno-English and Russian but not at all for the Spanish and Japanese listeners. Furthermore, in these subtests, the responses of the Japanese subjects were found to be very different from those of the European languages tested: there was a conspicuous gap in the cueing of *formal*, *stressed* and *relaxed* with any of the stimuli presented here for the Japanese listeners.

We would tend to conclude that these large differences in voice to affect mapping are likely to reflect underlying cross-cultural differences as they mostly involve attitudes and interpersonal stances, which are more likely to reflect culture-shaped learned behaviour. On the other hand, the possibility remains that the affective labels used in the listening tests are not fully equivalent for the different languages, and that this could have contributed to the results. This issue suggests some directions for future research where a richer range of affective states could be used to elicit affective response for stimuli such as the ones used here.

Another factor that could possibly contribute to the cross-language differences such as those found here and that requires further consideration is the neutral voice (baseline) in these different languages. If languages A and B have different neutral (baseline) voice and if the neutral voice in language A is similar to a quality which has affective colouring in language B, then it is plausible to hypothesise that these languages are likely to differ in voice-to- affect association. The study of such neutral (baseline) language-specific voice quality is another interesting direction for future research.

Further more subtle cross-language differences emerged that suggest that languages may differ in the relative sensitivity to voice quality or f_0 cues. There are indications that Japanese subjects may be more sensitive to f_0 cues where the European group of languages rely more heavily on voice quality cues.

It would be of great interest to explore further the areas of divergence that emerged in this study, in particular, those between Japanese on the one hand and the European languages on the other. A new study would need to take account of the potential limitations of the present study, and these are discussed in more detail below.

The synthesised stimuli The stimuli used in this experiment are necessarily a very limited subset of the infinite variety that occurs in real life. Certain qualities have been omitted, and the set of the 'VQ only' stimuli did not include extreme versions of voice qualities. The same points have to be made about the 'f_0 only' stimuli: although it was felt that a reasonable sampling of the differences in f_0 level and f_0 dynamics was represented, these are also a limited subset of the endless possibilities that exist. Similar holds for the combined 'VQ + f_0' stimuli: not all possible combinations of voice quality and f_0 contours have been covered in the present set.

More voice qualities will need to be included in the future work, such as for example harsh voice and falsetto. In the present study, harsh voice was omitted because it was not found to be very different from tense for Hiberno-English listeners in Gobl and Ní Chasaide (2003b). This limited the choices for the listeners in the present study, and of course might have been important in one of the other languages. A broader palette of voice qualities will be needed, particularly to explore those affects which did not achieve consistent affective signalling in this study, e.g., *relaxed-stressed*, *formal* (for Japanese subjects). Looking at production data from some Japanese corpora (Campbell et al., 2006) might be a good starting point .

A different approach to f_0 modelling is also potentially of interest for future work, including the possibility of varying the shape of the f_0 contour (e.g., so that the final accent is a rising rather than a falling one). The results of the present study suggest that the level of f_0 as well as the dynamics of the second peak (e.g., as in the f_0 'indignation' and f_0 'fear' contours) contribute significantly to the signalling of affect. In future work, a systematic manipulation of f_0 contour in terms of the global level, the relative height of the accented syllable peaks as well as the steepness of the declination line and the rate of change of the final fall (for H*L type of contours, for example) could be implemented in the generation of f_0 contours.

For both voice quality and pitch patterns, it would also be very interesting to look at qualities and intonation patterns from the production data in naturalistic corpora, and to use these to develop a more extended set of stimulus types which could be tested within the kind of perception experiment presented here. Judging by the results here, this would be particularly interesting to do for Japanese.

Stimuli evaluation In future work, the synthesised stimuli and the combinations of f_0 and voice quality would need to be tested in a preliminary test to assess their naturalness. It

would also be desirable that the raters' certainty in the way they assign affective ratings to a particular stimulus be assessed as part of the listening tests.

Affective labels and scales It would be of interest to see if the affective colouring of the voice quality stimuli could be elicited on a broader range of affective labels. This, for example, could include offering the listeners a rather narrow set of affective labels covering the 'families' of affective states as a first task, and if a certain voice quality does appear to impart affective colouring, give them a further task of providing a more specific label for the affective state (e.g., following the taxonomy of affective states in Baron-Cohen (2007)).

This study used a 7-point bipolar scale, which was found rather coarse and could allow for a fairly broad indication of affective colouring in a particular voice quality. It would be useful in future work to use a finer scale, for example, a unipolar 7-point scale. Another very different approach which might be particularly interesting would be to use the Feeltrace tool (Cowie et al., 2000) rather than a bipolar or unipolar rating scale.

The results of the cross-language study reported here suggested that achieving conceptual/semantic equivalence of the scale anchor terms in translation is crucial for the interpretation of results. It would be important in future work to adopt a stringent evaluating procedure of the semantic equivalence of the labels and to assess it in preliminary pilot perception tests.

7.2.2 Voice quality and loudness

The role of loudness variation (along with voice quality variation) in affect signalling has been explored in two related 'loudness experiments'. These experiments served first of all to demonstrate that the affective colourings associated with the different voice quality stimuli in the cross-language study (Chapter 4) were not simply due to the loudness differences among these stimuli. The second experiment explored further the contribution of loudness, as part of the voice quality, to the cueing of affect. The paradigm of the listening tests was the same as in the previous cross-language study, and the listeners were asked to judge the affective content of synthetic stimuli for the same set of affective labels.

Overall, the results of these experiments suggest that loudness variation on its own is rather ineffective for affect cueing while variation in voice quality is essential to the expression of affect. In Loudness experiment 1, stimuli incorporating voice quality variations consistently obtained higher ratings compared to the stimuli varying in loudness but with the

same modal voice quality. The results of the Loudness experiment 2 showed that equalising loudness of voice quality stimuli by setting their perceived loudness to match that of the modal voice does not have significant impact on their affective ratings: the voice quality stimuli appear capable of affect cueing for the majority of affects tested. However, loudness does contribute in affect signalling: when combined with tense-modal voice qualities, increased loudness can enhance signalling of high activation states such as *indignant, stressed* and *happy.*

7.2.3 Voice source parameters in affect portrayals: an illustrative analysis

A detailed voice source analysis was carried out on an utterance produced by a non-actor subject in ways that were intended to portray the following affective states: *bored, sad, happy, surprised, angry* as well as *neutral.* The salient voice source parameters were obtained through the inverse filtering of the utterance and the subsequent matching of the differentiated glottal flow waveform to the LF model to characterise the voice source in these portrayals. As the speech sample was not representative in a number of ways (portrayed rather than truly occurring affect; a single speaker; the analysis was carried out on only one utterance for each of the portrayed affects), the results are not presented to make any kind of definitive claims about how affects differ. Rather, this study was intended as an illustration of how a detailed parameter-rich analysis can proceed; it was also intended to explore in a preliminary way, the kinds of dimensions that may be important in affect differentiation. It is clear that an analysis of naturally occurring affect or affect obtained with a formal emotion induction procedure (e.g., the Velten technique) is something that needs to be done, and would certainly be of interest for future work.

This analytical fragment has served, however, to emphasise that a rich description of voice source variation is needed to capture the complexity of the differences that may occur. Although the global parameters such as RD and NAQ proposed in the literature are very useful, it is too early to assume that any single parameter can capture the subtlety of human voice variation in affect expression. The study further suggests that much of what we will need to understand concerns the dynamics of source variation. This also implies that future work will need to include the use of statistical methods that not only show the differences between the source parameters but that would allow one to get better insight into the dynamics of the voice source as well as the interaction among the parameters (e.g., regression analysis). The

affect-related source variation sits on a carrier wave which is itself dynamically varying, as part of the prosody of the utterance. A true understanding of how the voice expresses affect is likely to emerge when the underlying prosodic dynamics are studied in parallel. So a priority for future research will be to examine the lawful prosodic source variation, as the necessary baseline against which affect-related variation must be assessed.

References

Abe, I. (1998). Intonation in Japanese. In D. Hirst & A. Di Cristo (Eds.), *Intonation Systems* (pp. 360-375). Cambridge: Cambridge University Press.

Abelin, Å. (2004). *Spanish and Swedish interpretations of Spanish and Swedish emotions - the influence of facial expressions.* Paper presented at the FONETIK 2004, the XVII Swedish Phonetics Conference, Stockholm University, Sweden.

Abelin, Å., & Allwood, J. (2000). *Cross linguistic interpretation of emotional prosody.* Paper presented at the ITRW on Speech and Emotion, Newcastle, Northern Ireland.

Airas, M., & Alku, P. (2004). Emotions in short vowel segments: effects of the glottal flow as reflected by the normalised amplitude quotient. In *Affective Dialogue Systems* (Vol. 3068/2004, pp. 13-24). Berlin Heidelberg: Springer-Verlag

Airas, M., Pulakka, H., Bäckström, T., & Alku, P. (2005). *A toolkit for voice inverse filtering and parameterisation.* Paper presented at the Interspeech 2005, Lisbon, Portugal.

Albas, D. C., McCluskey, K. W., & Albas, C. A. (1976). Perception of the emotional content of speech: a comparison of two Canadian groups. *Journal of Cross-Cultural Psychology, 7*(4), 481-490.

Alcoba, S., & Murillo, J. (1998). Intonation in Spanish. In D. Hirst & A. Di Cristo (Eds.), *Intonation Systems* (pp. 152-166). Cambridge: Cambridge University Press.

Alku, P. (1992). Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering. *Speech Communication, 11*(2-3), 109-118.

Alku, P., Bäckström, T., & Vilkman, E. (2002). Normalized amplitude quotient for parameterization of the glottal flow. *The Journal of the Acoustical Society of America, 112*(2), 701-710.

Alku, P., Strik, H., & Vilkman, E. (1997). Parabolic spectral parameter - a new method for quantification of the glottal flow. *Speech Communication, 22*(1), 67-79.

Alku, P., & Vilkman, E. (1996). Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering. *Speech Communication, 18*(2), 131-138.

Alku, P., Vintturi, J., & Vilkman, E. (2001a). Evidence of the significance of secondary excitations of the vocal tract for vocal intensity. *Folia Phoniatrica et Logopaedica, 53*(4), 185-197.

Alku, P., Vintturi, J., & Vilkman, E. (2001b). Measuring the effect of fundamental frequency raising as a strategy for increasing vocal intensity in soft, normal and loud phonation. *Speech Communication, 38*, 321-334.

Andersen, P. A., & Guerrero, L. K. (Eds.). (1998). *Handbook of Communication and Emotion*. New York: Academic Press.

Andreeva, A. M., & Krivnova, O. F. (2005). *Laryngealisation in evaluatives in Russian dialogue speech (in Russian).* Paper presented at the Dialogue 2005, http://www.dialog-21.ru/Archive/2005.

Aubergé, V., & Cathiard, M. (2003). Can we hear the prosody of smile? *Speech Communication, 40*, 87-97.

Audibert, N., & Aubergé, V. (2007). Gradient or contours cues? A gating experiment for the timing of the emotional information. In *Affective Computing and Intelligent Interaction* (Vol. 4738, pp. 755-756): Springer Berlin / Heidelberg.

Audibert, N., Aubergé, V., & Rilliard, A. (2005). The relative weights of the different prosodic dimensions in expressive speech: a resynthesis study. In J. Tao, T. Tan & R. W. Picard (Eds.), *Affective Computing and Intelligent Interaction 2005* (pp. 527-534). Berlin Heidelberg: Springer-Verlag.

Audibert, N., Aubergé, V., & Rilliard, A. (2008). *How we are not equally competent for discriminating acted from spontaneous expressive speech.* Paper presented at the Speech Prosody 2008, Campinas, Brazil.

Bachorowski, J.-A. (1999). Vocal expression and perception of emotion. *Current Directions in Psychological Science, 8*(2), 53-57.

Bachorowski, J.-A., & Owren, M. (1995). Vocal expression of emotion: acoustic properties of speech are associated with emotional intensity and context. *Psychological Science, 6*(4), 210-224.

Bachorowski, J.-A., & Owren, M. J. (2003). Sounds of emotions. *Annals of the New York Academy of Sciences, 1000, Emotions Inside Out: 130 years after Darwin's The Expression of Emotions in Man and Animals*, 244-265.

Bachorowski, J.-A., & Owren, M. J. (2008). Emotion-related vocal acoustics. Cue-configuration, dimensional, and affect-induction perspectives. In K. Izdebski (Ed.), *Emotions in the Human Voice* (Vol. 1, pp. 87-99). San Diego, CA: Plural Publishing.

Bäckström, T., Alku, P., & Vilkman, E. (2002). Time-domain parameterization of the closing phase of glottal airflow waveform from voices over a large intensity range. *IEEE Transactions on Speech and Audio Processing, 10*(3), 186-192.

Baken, R. J., & Orlikoff, R. F. (2000). *Clinical Measurement of Speech and Voice* (2 ed.): Thomson Learning.

Balkwill, L.-L., Cuddy, L. L., & Thompson, W. F. (2004). *Acoustic cues in emotive music and speech across cultures.* Paper presented at the 8th International Conference on Music Perception and Cognition, Evanston, IL, USA.

Balkwill, L.-L., & Thompson, W. F. (1999). A cross-cultural investigation of the perception of emotion in music: psychophysiological and cultural cues. *Music Perception, 17*(1), 43-64.

Balkwill, L.-L., Thompson, W. F., & Matsunaga, R. (2004). Recognition of emotion in Japanese, Western, and Hindustani music by Japanese listeners. *Japanese Psychological Research, 46*(4), 337-349.

Ball, M. J., Esling, J. H., & Dickson, C. (2000). The transcription of voice quality. In R. D. Kent & M. J. Ball (Eds.), *Voice Quality Measurement* (pp. 49-58). San Diego: Singular Publishing Group.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology, 70*(3), 614-636.

Bänziger, T. (2004). *Communication vocale des émotions. Perception de l'expression vocale et attributions émotionnelles.* Unpublished PhD thesis, Geneva.

Bänziger, T., Morel, M., & Scherer, K. R. (2003). *Is there an emotion signature in intonational patterns? And can it be used in synthesis?* Paper presented at the Eurospeech 2003, Geneva, Switzerland.

Bänziger, T., & Scherer, K. R. (2003). *A study of perceived vocal features in emotional speech.* Paper presented at the VOQUAL'03, Geneva, Switzerland.

Bänziger, T., & Scherer, K. R. (2005). The role of intonation in emotional expression. *Speech Communication, 46*, 252-267.

Bänziger, T., & Scherer, K. R. (2007). Using actor portrayals to systematically study multimodal emotion expression: the GEMEP corpus. In A. Paiva & R. W. Picard (Eds.), *Affective Computing and Intelligent Interaction (ACII 2007)* (Vol. 4738, pp. 476-487). Berlin: Springer-Verlag.

Barkhuysen, P., Krahmer, E., & Swerts, M. (2007). *Cross-modal perception of emotional speech.* Paper presented at the XVIth International Congress of Phonetic Sciences, Saarbrücken, Germany.

Baron-Cohen, S. (2001a). Theory of mind and autism: a review. *The International Review of Research in Mental Retardation, 23*, 169-184.

Baron-Cohen, S. (2001b). Theory of mind in normal development and autism. *Prisme, 34*(174-183).

Baron-Cohen, S. (2007). Mind Reading: the Interactive Guide to Emotions. London: Jessica Kingsley Publishers.

Baron-Cohen, S., Riviere, A., Cross, P., Fukushima, M., Bryant, C., Sotillo, M., et al. (1996). Reading the mind in the face: a cross-cultural and developmental study. *Visual Cognition, 3*, 39-59.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The 'Reading the Mind in the Eyes' test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology & Psychiatry & Allied Disciplines, 42*(2), 241.

Baron-Cohen, S., Wheelwright, S., & Jolliffe, T. (1997). Is there a 'language of the eyes'? Evidence from normal adults, and adults with autism or Asperger syndrome. *Visual Cognition, 4*(3), 311-331.

Barrett, J., & Paus, T. (2002). Affect-induced changes in speech production. *Experimental Brain Research, 146*, 531-537.

Barrett, L. F. (2006). Are emotions natural kinds? *Perspectives on Psychological Science, 1*(1), 28-58.

Batliner, A., Fischer, K., Huber, R., Spilker, J., & Nöth, E. (2000). *Desperately seeking emotions or: actors, wizards and human beings.* Paper presented at the ISCA Workshop on Speech and Emotion, Belfast, Northern Ireland.

Batliner, A., Fischer, K., Huber, R., Spilker, J., & Nöth, E. (2003). How to find trouble in communication. *Speech Communication, 40*(1-2), 117-143.

Batliner, A., Schuller, B., Schaeffler, S., & Steidl, S. (2008). *Mothers, adults, children, pets - towards the acoustics of intimacy.* Paper presented at the International Conference on Acoustics, Speech and Signal Processing.

Batliner, A., Steidl, S., & Nöth, E. (2008). *Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus.* Paper presented at the LREC 2008, Marrakech, Morocco.

Beck, J. M. (2005). Perceptual analysis of voice quality: the place of vocal profile analysis. In W. J. Hardcastle & J. M. Beck (Eds.), *A Figure of Speech. A Festschrift for John Laver* (pp. 285-322). London: Lawrence Erlbaum Associates, Publishers.

Beckman, M. E. (1986). *Stress and Non-Stress Accent*: Mouton de Gruyter.

Behlau, M., & Gasparini, G. (2008). Vocal psychodynamics in the voice clinic. In K. Izdebski (Ed.), *Emotions in the Human Voice* (Vol. 2, pp. 273-281). San Diego, CA: Plural Publishing.

Behlau, M., & Gasparini, G. (2009). Preserving vocal emotions while dubbing into Brazilian Portugese. In K. Izdebski (Ed.), *Emotions in the Human Voice* (Vol. 3, pp. 277-287). San Diego, CA: Plural Publishing.

Beier, E. G., & Zautra, A. J. (1972). Identification of vocal communication of emotions across cultures. *Journal of Consulting and Clinical Psychology, 39*(1), 166-166.

Bele, I. V. (2005). Reliability in perceptual analysis of voice quality. *Journal of Voice, 19*(4), 555-573.

Biadsy, F., Rosenberg, A., Carlson, R., Hirshberg, J., & Strangert, E. (2008). *A cross-cultural comparison of American, Palestinian, and Swedish perception of charismatic speech.* Paper presented at the Speech Prosody 2008, Campinas, Brazil.

Boula de Mareüil, P., Célérier, P., & Toen, J. (2002). *Generation of emotions by a morphing technique in English, French and Spanish.* Paper presented at the Speech Prosody 2002, Aix-en-Provence, France.

Brandt, J. F., Ruder, K. F., & Shipp Jr, T. (1969). Vocal loudness and effort in continuous speech. *The Journal of the Acoustical Society of America, 46*(6B), 1543-1548.

Breitenstein, C., van Lancker, D., & Daum, I. (2001). The contribution of speech rate and pitch variation to the perception of vocal emotions in a German and an American sample. *Cognition and Emotion, 15*(1), 57-79.

Brislin, R. W. (1980). Translation and content analysis of oral and written material. In H. C. Triandis & J. W. Berry (Eds.), *Handbook of Cross-Cultural Psychology* (Vol. 2, pp. 389-444). Boston: Allyn and Bacon, Inc.

Bruyninckx, M., Harmegnies, B., Llisterri, J., & Poch-Olivé, D. (1994). Language-induced voice quality variability in bilinguals. *Journal of Phonetics, 22*, 19-31.

Bulut, M., Narayanan, S., & Johnson, L. (2004). Synthesising expressive speech: overview, challenges and open questions. In S. Narayanan & A. Alwan (Eds.), *Text to Speech Synthesis: New Paradigms and Anvances* (pp. 175-198): Prentice Hall.

Burgoon, J. K., & Hoobler, G. D. (2002). Nonverbal signals. In M. L. Knapp & J. A. Daly (Eds.), *Handbook of Interpersonal Communication* (3 ed., pp. 240-299): Sage Publishing, Inc.

Burkhardt, F. (2009). *Rule-based voice quality variation with formant synthesis.* Paper presented at the Interspeech 2009.

Burkhardt, F., Audibert, N., Malatesta, L., Türk, O., Arslan, L., & Aubergé, V. (2006). *Emotional prosody - does culture make a difference?* Paper presented at the Speech Prosody 2006, Dresden, Germany.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., & Weiss, B. (2005). *A database of German emotional speech.* Paper presented at the Interspeech 2005, Lisbon, Portugal.

Burkhardt, F., & Sendlmeier, W. F. (2000). *Verification of acoustical correlates of emotional speech using formant-synthesis.* Paper presented at the ITRW on Speech and Emotion, Newcastle, Northern Ireland.

Burkhardt, F., & Stegmann, J. (2009). *Emotional speech synthesis: applications, history and possible future.* Paper presented at the Elektronische Sprachsignalverarbeitung ESSV 2009.

Burleson, B. R., & Goldsmith, D. J. (1998). How the comforting process works: alleviating emotional distress through conversationally induced reappraisals. In P. A. Andersen & L. K. Gerrero (Eds.), *Handbook of Communication and Emotion.* New York: Academic Press.

Busso, C., & Narayanan, S. S. (2008). *Recording audio-visual emotional databases from actors: a closer look.* Paper presented at the 6th Language Resources and Evaluation Conference (LREC 2008), Marrakech, Morocco.

Buus, S., Florentine, M., & Poulsen, T. (1997). Temporal integration of loudness, loudness discrimination, and the form of the loudness function. *The Journal of the Acoustical Society of America, 101*(2), 669-680.

Cabral, J. P., & Oliveira, L. C. (2006). *EmoVoice: a system to generate emotion in speech.* Paper presented at the Interspeech 2006 - ICSLP, Pittsburgh, Pennsylvania, USA.

Campbell, N. (1995). *Loudness, spectral tilt and perceived prominence in dialogues.* Paper presented at the XIIIth International Congress of Phonetic Sciences, Stockholm, Sweden.

Campbell, N. (2004a). *Accounting for voice quality variation.* Paper presented at the Speech Prosody 2004, Nara, Japan.

Campbell, N. (2004b). *Listening between the lines: a study of paralinguistic information carried by tone-of-voice.* Paper presented at the International Symposium on Tonal Aspects of Languages: with Emphasis on Tone Languages, Beijing, China.

Campbell, N. (2004c). *Perception of affect in speech - towards an automatic processing of paralinguistic information in spoken conversation.* Paper presented at the Interspeech 2004 - ICSLP, Jeju Island, Korea.

Campbell, N. (2005). Developments in corpus-based speech synthesis: approaching natural conversational speech. *IEICE Transactions on Information and Systems, E88-D*(3), 376-383.

Campbell, N. (2008). Expressive/Affective Speech Synthesis. In J. Benesty, M. M. Sondhi & Y. Huang (Eds.), *Springer Handbook of Speech Processing* (pp. 505-517). Berlin Heidelberg: Springer-Verlag.

Campbell, N., & Mokhtari, P. (2003). *Voice quality: the 4th prosodic dimension.* Paper presented at the XVth International Congress of Phonetic Sciences, Barcelona, Spain.

Campbell, N., Sadanobu, T., Imura, M., Iwahashi, N., Noriko, S., & Douxchamps, D. (2006). *A multimedia database of meetings and informal interactions for tracking participant involvement and discourse flow.* Paper presented at the LREC 2006, Genoa, Italy.

Caridakis, G., Malatesta, L., Kessous, L., Amir, N., Raouzaiou, A., & Karpouzis, K. (2006). *Modelling naturalistic affective states via facial and vocal expression recognition.* Paper presented at the 8th International Conference on Multimidal Interfaces (ICMI 2006), Banff, Alberta, Canada.

Carlson, R., Granström, B., & Nord, L. (1992). *Experiments with emotive speech - acted utterances and synthesized replicas.* Paper presented at the 2nd International Conference on Spoken Language Processing (ICSLP 92), Banff, Alberta, Canada.

Cheang, H. S., & Pell, M. D. (2008). The sound of sarcasm. *Speech Communication, 50*(5), 366-381.

Childers, D. G., & Lee, C. K. (1991). Vocal quality factors: analysis, synthesis and perception. *The Journal of the Acoustical Society of America, 90*, 2394-2410.

Cornelius, R. R. (2000). *Theoretical approaches to emotion.* Paper presented at the ITRW on Speech and Emotion, Newcastle, Northern Ireland.

Cowie, R., & Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication, 40*, 5-32.

Cowie, R., Douglas-Cowie, E., Apolloni, B., Taylor, J., Romano, A., & Fellenz, W. (1999). What a neural net needs to know about emotion words. In N. Mastorakis (Ed.), *Computational Intelligence and Applications* (pp. 5311-5316).

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M. (2000). *'FEELTRACE': an instrument for recording perceived emotion in real time.* Paper presented at the ITRW on Speech and Emotion, Newcastle, Northern Ireland.

Crosswhite, K. (2003). *Spectral tilt as a cue to word stress in Polish, Macedonian, and Bulgarian.* Paper presented at the XVth International Congress of Phonetic Sciences, Barcelona, Spain.

Cullen, C., Vaughan, B., Kousidis, S., & McAuley, J. (2008). *Emotional speech corpus construction, annotation and distribution.* Paper presented at the LREC 2008, Marrakech, Morocco.

Cummings, K. E., & Clements, M. A. (1995a). Analysis of the glottal excitation of emotionally styled and stressed speech. *The Journal of the Acoustical Society of America, 98*, 88-98.

Cummings, K. E., & Clements, M. A. (1995b). Glottal models for digital speech processing: a historical survey and new results. *Digital Signal Processing, 5*, 21-42.

Daly, E. M., Lancee, W. J., & Polivy, J. (1983). A conical model for the taxonomy of emotional experience. *Journal of Personality and Social Psychology, 45*(2), 443-457.

Darwin, C. (1872/1999). *The Expression of the Emotions in Man and Animals*. London: Fontana Press.

Davitz, J. R. (1969). *The Language of Emotions*. New York: McGraw-Hill.

de Gelder, B., & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition and Emotion, 14*(3), 289-311.

Denes, P. B., & Pinson, E. N. (1993). *The Speech Chain: the Physics and Biology of Spoken Language*: W. H. Freeman & Co Ltd.

Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P. (2003). Emotional speech: towards a new generations of databases. *Speech Communication, 40*, 33-60.

Douglas-Cowie, E., & Cowie, R. *The vocal correlates of stress: a summary* http://www.image.ntua.gr/oresteia/private/contributed_docs/reports/Voice-and-Stress.pdf.

Douglas-Cowie, E., Cowie, R., Cox, C., Amir, N., & Heylen, D. (2008). *The Sensitive Artificial Listener: an induction technique for generating emotionally coloured conversation.* Paper presented at the LREC 2008, Marrakech, Morocco.

Douglas-Cowie, E., Cowie, R., Martin, J.-C., Devillers, L., & Cox, C. (2006). *HUMAINE Human Machine Interaction network on Emotions Mid Term Report on Database Exemplar Progress.*

Douglas-Cowie, E., Cowie, R., & Schröder, M. (2003). *The description of naturally occurring emotional speech.* Paper presented at the XVth International Congress of Phonetic Sciences, Barcelona, Spain.

Douglas-Cowie, E., Devillers, L., Martin, J.-C., Cowie, R., Savvidou, S., Abrilian, S., et al. (2005). *Multimodal databases of everyday emotion: facing up to complexity.* Paper presented at the Interspeech 2005.

Drioli, C., Tisato, G., Cosi, P., & Tesser, F. (2003). *Emotions and voice quality: experiments with sinusoidal modelling.* Paper presented at the VOQUAL'03.

Ekman, P. (1992). Are there basic emotions? *Psychological Review, 99*(3), 550-553.

Ekman, P. (1993). Facial expression and emotion. *American Psychologist, 48*(4), 384-392.

Ekman, P. (1999). Basic emotions. In T. Dalgleish & M. Power (Eds.), *Handbook of Cognition and Emotion.* Sussex, UK: John Wiley & Sons, Ltd.

Ekman, P. (Ed.). (1982). *Emotion in the Human Face.* Cambridge: Cambridge University Press.

Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., et al. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology, 53*(4), 712-717.

Elfenbein, H. A., & Ambady, N. (2002a). Is there an in-group advantage in emotion recognition? *Psychological Bulletin, 128*(2), 243-249.

Elfenbein, H. A., & Ambady, N. (2002b). On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological Bulletin, 128*(2), 203-235.

Elfenbein, H. A., & Ambady, N. (2003). Universals and cultural differences in recognizing emotions. *Current Directions in Psychological Science, 12*(5), 159-164.

Elfenbein, H. A., Ambady, N., Mandal, M. K., Harizuka, S., & Kumar, S. (2002). Brief report. Cross-cultural patterns on emotion recognition: highlighting design and analytical techniques. *Emotion, 2*(1), 75-84.

Ellsworth, P. C. (1991). Some implications of cognitive appraisal theories of emotion. In K. Strongman (Ed.), *International Review of Studies on Emotion.* New York: Wiley.

Enos, F., & Hirshberg, J. (2006). *A framework for eliciting emotional speech: capitalizing on the actor's process.* Paper presented at the LREC 2006 Workshop on Corpora for Research on Emotion and Affect, Genova, Italy.

Epstein, M. (2002). *Voice Quality and Prosody in English.* Unpublished PhD thesis.

Epstein, M. (2003). *Voice quality and prosody in English.* Paper presented at the XVth International Congress of Phonetic Sciences, Barcelona, Spain.

Erickson, D. (2005). Expressive speech: production, perception and application to speech synthesis. *Acoustical Science and Technology, 26*(4), 317-325.

Esling, J. H. (2000). Crosslinguistic aspects of voice quality. In R. D. Kent & M. J. Ball (Eds.), *Voice Quality Measurement* (pp. 25-35). San Diego: Singular Publishing Group.

Esling, J. H., & Wong, R. F. (1983). Voice quality settings and the teaching of pronunciation. *TESOL Quarterly, 17*(3), 89-95.

Fairbanks, G., & Pronovost, W. (1939). An experimental study of the pitch characteristics of the voice during the expression of emotion. *Communication Monographs, 6*(1), 87-104.

Fant, G. (1960/1971). *Acoustic theory of speech production* (2 ed.). The Hague: Mouton de Gruyter.

Fant, G. (1995). The LF-model revisited: transformations and frequency domain analysis. *STL-QPSR 2-3*, 119-156.

Fant, G. (1997). The voice source in connected speech. *Speech Communication, 22*, 125-139.

Fant, G., Kruckenberg, A., & Liljencrants, J. (2000). Acoustic-phonetic analysis of prominence in Swedish. In A. Botinis (Ed.), *Intonation. Analysis, Modelling and Technology* (pp. 55-86): Kluwer Academic Publishers.

Fant, G., Kruckenberg, A., Liljencrants, J., & Hertegård, S. (2000). Acoustic-phonetic studies of prominence in Swedish. *HTM-QPSR, 41*(2-3), 1-52.

Fant, G., Liljencrants, J., & Lin, Q. (1985). A four-parameter model of glottal flow. *STL-QPSR 4*, 1-13.

Fehr, B., & Russell, J. A. (1984). Concept of emotion viewed from a prototype perspective. *Journal of Experimental Psychology: General, 113*(3), 464-486.

Fitness, J., & Fletcher, G. J. O. (1993). Love, hate, anger, and jealousy in close relationships: A prototype and cognitive appraisal analysis. *Journal of Personality and Social Psychology, 65*(5), 942-958.

Fónagy, I. (2001). *Languages within Language: an Evolutive Approach*: John Benjamin Publishing Company.

Fónagy, I., & Magdics, K. (1963). Emotional patterns in intonation and music. *Zeitschrift für Phonetik, 16*, 293-326.

Frick, R. W. (1985). Communicating emotion: the role of prosodic features. *Psychological Bulletin, 97*(3), 412-429.

Frijda, N. H. (2005). Emotion experience. *Cognition and Emotion, 19*(4), 473-497.

Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *The Journal of the Acoustical Society of America, 27*(4), 765-768.

Fujisaki, H., & Ljungqvist, M. (1986). *Proposal and evaluation of models for the glottal source waveform.* Paper presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '86).

Gannon, K. M., & Ostrom, T. M. (1996). How meaning is given to rating scales: the effects of response language on category activation. *Journal of Experimental Social Psychology, 32*, 337-360.

Gerratt, B. R., & Kreiman, J. (2000). Theoretical and methodological development in the study of pathological voice quality. *Journal of Phonetics, 28*(335-432).

Gerratt, B. R., & Kreiman, J. (2001). Measuring vocal quality with speech synthesis. *The Journal of the Acoustical Society of America, 110*(5), 2560-2566.

Glave, R. D., & Rietveld, A. C. M. (1975). Is the effort dependence of speech loudness explicable on the basis of acoustical cues? *The Journal of the Acoustical Society of America, 58*(4), 875-879.

Glave, R. D., & Rietveld, A. C. M. (1979). Bimodal cues for speech loudness. *The Journal of the Acoustical Society of America, 66*(4), 1018-1022.

Gobl, C. (1988). Voice source dynamics in connected speech. *STL-QPSR 1*, 123-159.

Gobl, C. (1989). A preliminary study of acoustic voice quality correlates. *STL-QPSR 30*(4), 9-22.

Gobl, C. (2003). *The Voice Source in Speech Communication. Doctoral thesis*. Stockholm: KTH, Department of Speech, Music and Hearing.

Gobl, C., Bennett, E., & Ní Chasaide, A. (2002). *Expressive synthesis: how crucial is voice quality?* Paper presented at the IEEE Workshop on Speech Synthesis, Santa Monica, California, USA.

Gobl, C., & Ní Chasaide, A. (1992). Acoustic characteristics of voice quality. *Speech Communication, 11*, 481-490.

Gobl, C., & Ní Chasaide, A. (1999a). Techniques for investigating laryngeal articulation (Section B: Techniques for analysing the voice source). In W. J. Hardcastle & N. Hewlett (Eds.), *Coarticulation: Theory, Data and Techniques* (pp. 300-321). Cambridge: Cambridge University Press.

Gobl, C., & Ní Chasaide, A. (1999b). Voice source variation in the vowel as a function of consonantal context. In W. J. Hardcastle & N. Hewlett (Eds.), *Coarticulation: Theory, Data and Techniques* (pp. 122-143). Cambridge: Cambridge University Press.

Gobl, C., & Ní Chasaide, A. (2000). *Testing affective correlates of voice quality through analysis and resynthesis.* Paper presented at the ITRW on Speech and Emotion, Newcastle, Northern Ireland.

Gobl, C., & Ní Chasaide, A. (2002). Dynamics of the glottal source signal. In E. Keller, G. Bailly, A. Monaghan, J. Terken & M. Huckvale (Eds.), *Improvements in Speech Synthesis* (pp. 273-283): John Wiley & Sons, Ltd.

Gobl, C., & Ní Chasaide, A. (2003a). *Amplitude-based source parameters for measuring voice quality.* Paper presented at the VOQUAL'03, Geneva, Switzerland.

Gobl, C., & Ní Chasaide, A. (2003b). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication, 40*(1-2), 189-212.

Gordeeva, O., Mennen, I., & Scobbie, J. (2003). *Vowel duration and spectral balance in Scottish English and Russian.* Paper presented at the XVth International Congress of Phonetics Sciences, Barcelona, Spain.

Gordon, M., & Ladefoged, P. (2001). Phonation types: a cross-linguistic overview. *Journal of Phonetics, 29*, 383-406.

Graham, C. R., Hamblin, A. W., & Feldstein, S. (2001). Recognition of emotion in English by speakers of Japanese, Spanish and English. *IRAL - International Review of Applied Linguistics in Language Teaching, 39*(1), 19-37.

Grandjean, D., Bänziger, T., & Scherer, K. R. (2006). Intonation as an interface between language and affect. *Progress in Brain Research, 156*, 235-268.

Granqvist, S. (1996). Enhancements to the visual analogue scale, VAS, for listening tests. *TMH-QPSR 37*(4), 61-65.

Greasly, P., Sherrard, C., & Waterman, M. (2000). Emotion in language and speech: methodological issues in naturalistic approaches. *Language and Speech, 43*, 355-375.

Grichkovtsova, I., & Mennen, I. (2008). Child speech and emotions: a cross-linguistic perspective. In K. Izdebski (Ed.), *Emotions in the Human Voice* (Vol. 2, pp. 35-47). San Diego, CA: Plural Publishing.

Grimm, M., Kroschel, K., Mower, E., & Narayanan, S. (2007). Primitives-based evaluation and estimation of emotions in speech. *Speech Communication, 49*(10-11), 787-800.

Guerrero, L. K., Andersen, P., & Trost, M. R. (1998). Communication and emotion: basic concepts and approaches. In P. A. Andersen & L. K. Guerrero (Eds.), *Handbook of Communication and Emotion.* New York: Academic Press.

Hager, J. C., & Ekman, P. (1983). The inner and outer meanings of facial expressions. In *Social Psychophysiology: a Sourcebook.* New York: Guilford Press.

Hall, E. (1976). *Beyond Culture*. New York: Anchor Press.

Hammerschmidt, K., & Jürgens, U. (2007). Acoustical correlates of affective prosody. *Journal of Voice, 21*(5), 531-540.

Hanson, H. M. (1997). Glottal characteristics of female speakers: acoustic correlates. *The Journal of the Acoustical Society of America, 101*(1), 466-481.

Hanson, H. M., & Chuang, E. S. (1999). Glottal characteristics of male speakers: acoustic correlates and comparison with female data. *The Journal of the Acoustical Society of America 106*(2), 1064-1077.

Hatta, T., & Nachshon, I. (1988). Ear differences in evaluating emotional overtones of unfamiliar speech by Japanese and Israelis. *International Journal of Psychology, 23*(3), 293-302.

Heise, D. R. (1970). The semantic differential and attitude research. In G. F. Summers (Ed.), *Attitude Measurement* (pp. 235-253). Chicago: Rand McNally.

Heldner, M. (2003). On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish. *Journal of Phonetics, 31*, 39-62.

Henrich, N. (2006). Mirroring the voice from Garcia to the present day: some insight into singing voice registers. *Logopedics Phoniatrics Vocology, 31*(1), 3-14.

Hirano, M. (1981). *Clinical Examination of Voice*. New York: Springer Verlag.

Honda, K., Maeda, S., & Sugito, M. (2007). *Articulation changes in different voicing patterns.* Paper presented at the XVIth International Congress of Phonetic Sciences, Saarbrücken, Germany.

http://www.ncvs.org/ncvs/tutorials/voiceprod/tutorial/quality.html.

Ide, S., & Yoshida, M. (1999). Sociolinguistics: honorifics and gender differences. In N. Tsujimura (Ed.), *The Handbook of Japanese Linguistics* (pp. 444-480): Blackwell Publishing.

Iseli, M., Shue, Y.-L., & Alwan, A. (2007). Age, sex, and vowel dependencies of acoustic measures related to the voice source. *The Journal of the Acoustical Society of America, 121*(4), 2283-2295.

Iseli, M., Shue, Y.-L., Epstein, M. A., Keating, P., Kreiman, J., & Alwan, A. (2006). *Voice source correlates of prosodic features in American English.* Paper presented at the Interspeech 2006 - ICSLP, Pittsburgh, Pennsylvania, USA.

Ishi, C. T., Ishiguro, H., & Hagita, N. (2008). *The role of breathy/whispery voice qualities in dialogue speech.* Paper presented at the Speech Prosody 2008, Campinas, Brazil.

ISO. (1975). Acoustics. Methods for calculating loudness level. International standard ISO 532: International Organisation for Standardisation.

Ito, M. (2005). *The contribution of voice quality to the expression of politeness: an experimental study.* Unpublished PhD thesis, University of Edinburgh, Edinburgh.

Izard, C. E. (1980). Cross-cultural perspectives on emotion and emotion communication. In H. C. Triandis & W. Lonner (Eds.), *Handbook of Cross-Cultural Psychology* (Vol. 3, pp. 185-221). Boston: Allyn and Bacon, Inc.

Izard, C. E. (1990). Facial expressions and the regulation of emotions. *Journal of Personality and Social Psychology, 58*(3), 487-498.

Izard, C. E. (1994). Innate and universal facial expressions: evidence from developmental and cross-cultural research. *Psychological Bulletin, 115*(2), 288-299.

Izdebski, K. (2008). Pathologic phonation connotes and evokes wrong emotive reaction in listeners. In K. Izdebski (Ed.), *Emotions in the Human Voice* (Vol. 2, pp. 119-127). San Diego, CA: Plural Publishing.

Johnson-Laird, P. N., & Oatley, K. (1989). The language of emotions: an analysis of a semantic field. *Cognition and Emotion, 3*(2), 81-123.

Johnstone, T. (1996). *Emotional speech elicited using computer games.* Paper presented at the 4th International Conference on Spoken Language Processing, Philadelphia, USA.

Johnstone, T., Banse, R., & Scherer, K. R. (1995). *Acoustic profiles in prototypical vocal expressions of emotions.* Paper presented at the XIIIth International Congress of Phonetic Sciences, Stockholm, Sweden.

Johnstone, T., & Scherer, K. R. (1999). *The effects of emotions on voice quality.* Paper presented at the XIVth International Congress of Phonetic Sciences, San Francisco.

Johnstone, T., & Scherer, K. R. (2000). Vocal communication of emotion. In M. Lewis & J. M. Haviland-Jones (Eds.), *Handbook of Emotions* (2 ed., pp. 220-235). New York: Guilford Press.

Johnstone, T., Van Reekum, C. M., Bänziger, T., Hird, K., Kirsner, K., & Scherer, K. R. (2007). The effect of difficulty and gain versus loss on vocal physiology and acoustics. *Psychophysiology, 44*, 827-837.

Johnstone, T., van Reekum, C. M., Hird, K., Kirsner, K., & Scherer, K. R. (2005). Affective speech elicited with a computer game. *Emotion, 5*(4), 513-518.

Johnstone, T., van Reekum, C. M., & Scherer, K. R. (2001). Vocal expression correlates of appraisal processes. In K. R. Scherer, A. Schorr & T. Johnstone (Eds.), *Appraisal Processes in Emotion: Theory, Methods, Research* (pp. 271-284). Oxford: Oxford University Press.

Juslin, P. N., & Laukka, P. (2001). Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. *Emotion, 1*(4), 381-412.

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: different channels, same code? *Psychological Bulletin, 5*, 770-814.

Juslin, P. N., & Scherer, K. R. (2005). Vocal expression of affect. In J. Harrigan, R. Rosenthal & K. R. Scherer (Eds.), *The New Handbook of Methods in Nonverbal Behavior Research* (pp. 65-135). Oxford: Oxford University Press

Kane, J. C., & Gobl, C. (2009). *Automatic parameterisation of the glottal waveform combining time and frequency domain measures.* Paper presented at the MAVEBA 2009, Florence, Italy.

Kappas, A., Hess, U., & Scherer, K. R. (1991). Voice and emotion. In R. S. Feldman & B. Rimé (Eds.), *Fundamentals of Nonverbal Behavior* (pp. 200-238). Cambridge: Cambridge University Press.

Kappas, A., & Poliakova, N. (2008). Judgment of the affective valence of spontaneous vocalisations. In K. Izdebski (Ed.), *Emotions in the Human Voice* (Vol. 1, pp. 109-122). San Diego, CA: Plural Publishing.

Keating, P. A., & Esposito, C. (2007). Linguistic voice quality. *UCLA Working Papers in Phonetics, 105*, 85-91.

Keller, E. (2005). The analysis of voice quality in speech processing. In *Nonlinear Speech Modeling and Applications* (Vol. 3445/2005, pp. 54-73). Berlin / Heidelberg: Springer

Kitayama, S., & Ishii, K. (2002). Word and voice: spontaneous attention to emotional utterances in two languages. *Cognition and Emotion, 16*(1), 29-59.

Klasmeyer, G. (1997). *The perceptual importance of selected voice quality parameters.* Paper presented at the ICASSP'97, Munich, Germany.

Klasmeyer, G., & Sendlmeier, W. F. (1995). *Objective voice parameters to characterise the emotional content in speech.* Paper presented at the XIIIth International Congress of Phonetic Sciences, Stockholm.

Klasmeyer, G., & Sendlmeier, W. F. (2000). Voice and emotional states. In R. D. Kent & M. J. Ball (Eds.), *Voice Quality Measurement* (pp. 339-357). San Diego: Singular Publishing Group.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *The Journal of the Acoustical Society of America, 67*, 971-995.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America, 87*, 820-857.

Kleinginna, P. R., & Kleinginna, A. M. (1981). A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and Emotion, 5*(4), 345-379.

Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005). Loudness predicts prominence: fundamental frequency lends little. *The Journal of the Acoustical Society of America, 118*(2), 1038-1054.

Kramer, E. (1963). Judgment of personal characteristics and emotions from nonverbal properties of speech. *Psychological Bulletin, 60*(4), 408-420.

Kramer, E. (1964). Elimination of verbal cues in judgments of emotion from voice. *The Journal of Abnormal and Social Psychology, 68*(4), 390-396.

Kreiman, J., & Gerratt, B. R. (1998). Validity of rating scale measures of voice quality. *The Journal of the Acoustical Society of America, 104*(3), 1598-1608.

Kreiman, J., & Gerratt, B. R. (2000a). Measuring vocal quality. In R. D. Kent & M. J. Ball (Eds.), *Voice Quality Measurement* (pp. 73-101). San Diego: Singular Publishing Group.

Kreiman, J., & Gerratt, B. R. (2000b). Sources of listener disagreement in voice quality assessment. *The Journal of the Acoustical Society of America, 108*(2), 1867-1876.

Kreiman, J., Gerratt, B. R., & Antoñanzas-Barroso, N. (2007). Measures of the glottal source spectrum. *Journal of Speech, Language, and Hearing Research, 50*, 595-610.

Kreiman, J., Gerratt, B. R., & Ito, M. (2007). When and why listeners disagree in voice quality assessment tasks. *The Journal of the Acoustical Society of America, 122*(4), 2354-2364.

Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., & Berke, G. S. (1993). Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *Journal of Speech and Hearing Research, 36*, 21-40.

Kreiman, J., Vanlancker-Sidtis, D., & Gerratt, B. R. (2003). *Defining and measuring voice quality.* Paper presented at the VOQUAL'03, Geneva, Switzerland.

Kreiman, J., Vanlancker-Sidtis, D., & Gerratt, B. R. (2004). *Defining and measuring voice quality.* Paper presented at the From Sound to Sense, MIT.

Kreiman, J., Vanlancker-Sidtis, D., & Gerratt, B. R. (2005). Perception of voice quality. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 338-362): Blackwell Publishing.

Ladd, D. R., Scherer, K. R., & Silverman, K. (1986). An integrated approach to studying intonation and attitude. In K. Johns-Lewis (Ed.), *Intonation in Discourse* (pp. 125-138). London: Croom Helm.

Ladd, D. R., Silverman, K. E. A., Tolkmitt, F., Bergmann, G., & Scherer, K. R. (1985). Evidence for the independent function of intonation contour type, voice quality, and *F*0 range in signaling speaker affect *The Journal of the Acoustical Society of America, 78*(2), 435-444.

Ladefoged, P. (2003). *Phonetic Data Analysis: and Introduction to Fieldwork and Instrumental Techniques*: Plackwell Publishing.

Ladefoged, P., & Maddieson, I. (1996). *The Sounds of the World's Languages*. Oxford: Blackwell Publishers.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Larsen, R. J., & Diener, E. (1992). Promises and problems with the circumplex model of emotion. *Review of Personality and Social Psychology, 13*, 25-59.

Laukka, P. (2004). Vocal Expression of Emotion, *Acta Universitatis Upsaliensis. Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences* (Vol. 140). Uppsala: Uppsala University.

Laukka, P. (2008). Research on vocal expression of emotion: state of the art and future directions. In K. Izdebski (Ed.), *Emotions in the Human Voice* (Vol. 1, pp. 153-169). San Diego, CA: Plural Publishing.

Laukkanen, A.-M., Alku, P., Airas, M., & Waaramaa, T. (2008). The role of voice quality in the expression and perception of emotion. In K. Izdebski (Ed.), *Emotions in the Human Voice* (pp. 171-184). San Diego, CA: Plural Publishing.

Laukkanen, A.-M., Vilkman, E., Alku, P., & Oksanen, H. (1996). Physical variations related to stress and emotional state: a preliminary study. *Journal of Phonetics, 24*, 313-335.

Laver, J. (1980). *The Phonetic Description of Voice Quality*. Cambridge: Cambridge University Press.

Laver, J. (2000). Phonetic evaluation of voice quality. In R. D. Kent & M. J. Ball (Eds.), *Voice Quality Measurement* (pp. 37-48). San Diego: Singular Publishing Group.

Lazarus, R. S. (1991). *Emotion and Adaptation*. New York; Oxford: Oxford University Press.

Lehman, D. R., Chiu, C., & Schaller, M. (2004). Psychology and culture. *Annual Review of Psychology, 55*(1), 689-714.

Leinonen, L., Hiltunen, T., Linnankoski, I., & Laakso, M.-L. (1997). Expression of emotional-motivational connotations with a one-word utterance. *The Journal of the Acoustical Society of America, 102*(3), 1853-1863.

Levy, R. I. (1984). Emotion, knowing, and culture. In R. A. Shweder & R. A. LeVine (Eds.), *Culture Theory: Essays on Mind, Self, and Emotion*. Cambridge: Cambridge University Press.

Lieberman, P., & Michaels, S. B. (1962). Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech. *The Journal of the Acoustical Society of America, 34*(7), 922-927.

Linnankoski, I., Leinonen, L., Vihla, M., Laakso, M.-L., & Carlson, S. (2005). Conveyance of emotional connotations by a single word in English. *Speech Communication, 45*(1), 27-39.

Luce, R. D., & Green, D. M. (1972). A neural timing theory for response times and the psychophysics of intensity. *Psychological Review, 79*, 14-57.

Maekawa, K. (1998). *Phonetic and phonological characteristics of paralinguistic information in spoken Japanese.* Paper presented at the 5th International Conference on Spoken Language Processing - ICSLP'98, Sidney, Australia.

Mahshie, J., & Gobl, C. (1999). *Effects of varying LF parameters on KLSYN88 synthesis.* Paper presented at the XIVth International Congress of Phonetic Sciences, San Francisco.

Martin, M. (1990). On the induction of mood. *Clinical Psychology Review, 10*(6), 669-697.

Mathon, C., & de Abreu, S. (2007). Emotion from speakers to listeners: perception and prosodic characterization of affective speech. In C. Müller (Ed.), *Speaker Classification II* (Vol. 4441, pp. 70-82). Berlin Heidelberg: Springer-Verlag.

Mathon, C., de Abreu, S., & Perekopska, D. (2006). *Perception of anger in French as foreign language: experimental protocol and preliminary results.*

Matsumoto, D., & Assar, M. (1992). The effects of language on judgments of universal facial expressions of emotion. *Journal of Nonverbal Behaviour, 16*(2), 85-99.

McCluskey, K. W., & Albas, D. C. (1981). Perception of the emotional content of speech by Canadian and Mexican children, adolescents, and adults. *International Journal of Psychology, 16*(2), 119-132.

McCluskey, K. W., Albas, D. C., Niemi, R. R., Cuevas, C., & Ferrer, C. A. (1975). Cross-cultural differences in the perception of the emotional content of speech: a study of the development of sensitivity in Canadian and Mexican children. *Developmental Psychology, 11*(5), 551-555.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*(1), 30-46.

Mesquita, B. (2003). Emotions as dynamic social phenomena. In R. J. Davidson, K. R. Scherer & D. J. Goldsmith (Eds.), *Handbook of Affective Sciences* (pp. 871-890). Oxford: Oxford University Press.

Mesquita, B., & Ellsworth, P. C. (2001). The role of culture in appraisal. In K. R. Scherer, A. Schorr & T. Johnstone (Eds.), *Appraisal Processes in Emotion: theory, Methods, Research* (pp. 233-248). Oxford: Oxford University Press.

Mesquita, B., & Frijda, N. H. (1992). Cultural variations in emotions: a review. *Psychological Bulletin, 112*(2), 179-204.

Mesquita, B., Frijda, N. H., & Scherer, K. R. (1997). Culture and emotion. In J. W. Berry, P. R. Dasen & T. S. Saraswathi (Eds.), *Handbook of Cross-Cultural Psychology: Basic Processes and Human Development* (2 ed., Vol. 2): Longwood Professional Books.

Michaelis, D., Gramss, T., & Strube, H. W. (1997). Glottal-to-Noise Excitation ratio - a new measure for describing pathological voices. *Acustica, 83*, 700-706.

Möbius, B. (2003). *Gestalt psychology meets phonetics - an early experimental study of intrinsic F0 and intensity.* Paper presented at the XVth International Congress of Phonetic Sciences, Barcelona, Spain.

Monahan, P. (1996). *Systems for voice source analysis.* Unpublished PhD thesis, Trinity College Dublin, Dublin.

Montero, J. M., Gutiérrez Arriola, J. M., Colás, J., Enríquez, E., & Pardo, J. M. (1999). *Analysis and modelling of emotional speech in Spanish.* Paper presented at the XIVth International Congress of Phonetic Sciences, San Francisco, USA.

Montero Martínez, J. M., Gutiérrez Arriola, J. M., de Córdoba Herralde, R., Enríquez Carrasco, E. V., & Pardo Muñoz, J. M. (2002). The role of pitch and tempo in Spanish emotional speech. In E. Keller, G. Bailly, A. Monaghan, J. Terken & M. Huckvale (Eds.), *Improvements in Speech Synthesis* (pp. 276-251): John Wiley & Sons, Ltd.

Monzo, C., Alías, F., Iriondo, I., Gonzalvo, X., & Planet, S. (2007). *Discriminating expressive speech styles by voice quality parameterization.* Paper presented at the XVIth International Congress of Phonetic Sciences, Saarbrücken, Germany.

Moore, B. C. J. (2003). *An Introduction to the Psychology of Hearing* (5th ed.). London: Academic Press.

Moore II, E., Clements, M. A., Peifer, J. W., & Weisser, L. (2003). *Investigating the role of glottal features in classifying clinical depression.* Paper presented at the 25th Annual International Conference of the IEEE EMBS, Cancun, Mexico.

Moore II, E., Clements, M. A., Peifer, J. W., & Weisser, L. (2008). Critical analysis of the impact of glottal features in the classification of clinical depression. *Biomedical Engineering, 55*(1), 96-107.

Mozziconacci, S. (1995). *Pitch variations and emotions in speech.* Paper presented at the XIIIth International Congress of Phonetic Sciences, Stockholm.

Mozziconacci, S. (1998). *Speech Variability and Emotion: Production and Perception. Doctoral Thesis*. Eindhoven: Technische Universiteit Eindhoven.

Mozziconacci, S. (2002). *Prosody and emotions.* Paper presented at the Speech Prosody 2002, Aix-en-Provence, France.

Murphy, P. J., & Laukkanen, A.-M. (2009). Electroglottogram analysis of emotionally styled phonation. In *Multimodal Signals: Cognitive and Algorithmic Issues* (pp. 264-270): Springer Berlin/Heidelberg.

Murray, I. R., & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America, 93*, 1907–1108.

Murray, I. R., Arnott, J. L., & Rohwer, E. A. (1996). Emotional stress in synthetic speech: progress and future directions. *Speech Communication, 20*, 85-91.

Ní Chasaide, A., & Gobl, C. (1995). *Towards acoustic profiles of phonatory qualities.* Paper presented at the XIIIth International Congress of Phonetic Sciences, Stockholm.

Ní Chasaide, A., & Gobl, C. (1997). Voice source variation. In W. J. Hardcastle & J. Laver (Eds.), *The Handbook of Phonetic Sciences* (pp. 427-461). Oxford: Blackwell.

Ní Chasaide, A., & Gobl, C. (2002). Voice quality and the synthesis of affect. In E. Keller, G. Bailly, A. Monaghan, J. Terken & M. Huckvale (Eds.), *Improvements in Speech Synthesis* (pp. 252-263). New York: John Wiley & Sons.

Ní Chasaide, A., & Gobl, C. (2004a). *Decomposing linguistic and affective components of phonatory quality.* Paper presented at the Interspeech 2004, Jeju Island, Korea.

Ní Chasaide, A., & Gobl, C. (2004b). *Voice quality and f0 in prosody: towards a holistic account.* Paper presented at the Speech Prosody 2004, Nara, Japan.

Ní Chasaide, A., & Gobl, C. (2005). On the relation between phonatory quality and affect. In W. J. Hardcastle & J. M. Beck (Eds.), *A Figure of Speech: a Festschrift for John Laver.* London: Lawrence Erlbaum Associates, Publishers.

Ní Chasaide, A., Gobl, C., & Monahan, P. (1992). A technique for analysing voice quality in pathological and normal speech. *Journal of Clinical Speech and Language Studies, 2*, 1-16

Núñez Batailla, F., Corte Santos, P., Sequeiros Santiago, G., Señaris Gonzáles, B., & Suárez Nieto, C. (2004). Perceptual evaluation of dysphonia: correlation with acoustic parameters and reliability. *Acta Otorrinolaringológica Española, 55*, 282-287.

O'Connor, J. D., & Arnold, G. F. (1973). *Intonation of Colloquial English* (2 ed.). London: Longman.

Oatley, K., Keltner, D., & Jenkins, J. M. (2006). *Understanding Emotions* (2 ed.). Oxford: Blackwell Publishing Ltd.

Ofuka, E., McKeown, J. D., Waterman, M. G., & Roach, P. J. (2000). Prosodic cues for rated politeness in Japanese speech. *Speech Communication, 32*(3), 199-217.

Ogarkova, A., Borgeaud, P., & Scherer, K. R. (2009). Language and culture in emotion research: a multidisciplinary perspective. *Social Science Information, 48*(3), 339-357.

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of f0 of voice. *Phonetica, 41*, 1-16.

Ohala, J. J. (1994). The frequency code underlies the sound-symbolic use of voice pitch. In L. Hinton, J. Nichols & J. J. Ohala (Eds.), *Sound Symbolism* (pp. 325-347). Cambridge: Cambridge University Press.

Ohala, J. J. (1996). *Ethological theory and the expression of emotion in the voice.* Paper presented at the 4th International Conference on Spoken Language Processing, Philadelphia, USA.

Ortony, A., & Turner, T. (1990). What's basic about basic emotions? *Psychological Review, 97*, 315-331.

Osgood, C. E., May, W. H., & Miron, M. S. (1975). *Cross-Cultural Universals of Affective Meaning*. London: University of Illinois Press.

Paeschke, A. (2004). *Global trend of fundamental frequency in emotional speech.* Paper presented at the Speech Prosody 2004, Nara, Japan.

Paeschke, A., Kienast, M., & Sendlmeier, W. F. (1999). *F0-contours in emotional speech.* Paper presented at the XIVth International Congress of Phonetic Sciences, San Francisco, USA.

Pakosz, M. (1983). Attitudinal judgments in intonation: some evidence for a theory. *Journal of Psycholinguistic Research, 12*(3), 311-326.

Pierrehumbert, J. B. (1989). A preliminary study of the consequences of intonation for the voice source. *STL-QPSR, 30*(4), 23-36.

Pittam, J., & Scherer, K. R. (1993). Vocal expression and communication of emotion. In M. Lewis & J. M. Haviland (Eds.), *Handbook of Emotions* (pp. 185-197). New York: Guilford Press.

Plack, C. J., & Carlyon, R. P. (1995). Loudness perception and intensity coding. In B. C. J. Moore (Ed.), *Hearing* (pp. 123-160). New York: Academic Press.

Planalp, S. (1998). Communicating emotions in everyday life: cues, channels and processes. In P. A. Andersen & L. K. Guerrero (Eds.), *Handbook of Communication and Emotion*. New York: Academic Press.

Plutchik, R. (1980). *Emotion: a Psychoevolutionary Synthesis*. New York: Harper & Row.

Plutchik, R., & Kellerman, H. (Eds.). (1980). *Emotion: Theory, Research and Experience*. New York: Academic Press.

Porter, E. R., & Samovar, L. A. (1998). Cultural influences in emotional expression: implications for intercultural communication. In P. A. Andersen & L. K. Guerrero (Eds.), *Handbook of Communication and Emotion*. New York: Academic Press.

Prinz, J. J. (2004a). *Gut Reactions: a Perceptual Theory of Emotion*. Oxford: Oxford University Press.

Prinz, J. J. (2004b). Which emotions are basic? In D. Evans & P. Cruse (Eds.), *Emotion, Evolution and Rationality* (pp. 69-87). Oxford: Oxford University Press.

Protopapas, A., & Lieberman, P. (1997). Fundamental frequency of phonation and perceived emotional stress. *The Journal of the Acoustical Society of America, 101*(4), 2267-2277.

Richell, R. A., & Anderson, M. (2004). Reproducibility of negative mood induction: a self-referent plus musical mood induction procedure and a controllable/uncontrollable stress paradigm. *Journal of Psychopharmacology, 18*(1), 94-101.

Robinson, D. W., & Dadson, R. S. (1956). A redetermination of the equal loudness relations for pure tones. *British Journal of Applied Physics, 7*, 166-181.

Rossato, S., Audibert, N., & Aubergé, V. (2004). *Emotional voice measurement: a comparison of articulatory-EGG and acoustic-amplitude parameters.* Paper presented at the Speech Prosody 2004, Nara, Japan.

Rothenberg, M. (1973). A new inverse-filtering technique for deriving the glottal air flow waveform during voicing. *The Journal of the Acoustical Society of America, 53*(6), 1632-1645.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161-1178.

Russell, J. A. (1991). Culture and the categorisation of emotions. *Psychological Bulletin, 110*(3), 426-450.

Russell, J. A. (1994). Is there universal recognition of emotion from facial expressions? A review of the cross-cultural studies. *Psychological Bulletin, 115*(1), 102-141.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review, 110*(1), 145-172.

Russell, J. A., Bachorowski, J.-A., & Fernández-Dols, J.-M. (2003). Facial and vocal expression of emotion. *Annual Review of Psychology, 54*, 329-349.

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotions: dissecting the elephant. *Journal of Personality and Social Psychology, 76*(5), 805-819.

Russell, J. A., & Carroll, J. M. (1999). On the bipolarity of positive and negative affect. *Psychological Bulletin, 125*(1), 3-30.

Russell, J. A., & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality, 11*, 273-294.

Ryan, C., Ní Chasaide, A., & Gobl, C. (2003). *Voice quality variation and the perception of affect: continuous or categorical?* Paper presented at the XVth International Congress of Phonetic Sciences, Barcelona, Spain.

Sadanobu, T. (2004). A natural history of Japanese pressed voice. *Journal of the Phonetic Society of Japan, 8*(1), 29-44.

Sander, D., Grandjean, D., & Scherer, K. R. (2005). A systems approach to appraisal mechanisms in emotion. *Neural Networks, 18*, 317-352.

Sawamura, K., Dang, J., Akagi, M., Erickson, D., Li, A., Sakuraba, K., et al. (2007). *Common factors in emotion perception among different cultures.* Paper presented at the XVIth International Congress of Phonetic Sciences, Saarbrücken, Germany.

Scharf, B. (1978). Loudness. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of Perception* (Vol. 4. Hearing, pp. 187-242). New York: Academic Press.

Scherer, K. R. (1978). Presonality inference from voice quality: the loud voice of extroversion. *European Journal of Social Psychology, 8*, 467-487.

Scherer, K. R. (1982a). Emotion as a process: function, origin and regulation. *Social Science Information, 21*, 555-570.

Scherer, K. R. (1982b). Methods of research on vocal communication: paradigms and parameters. In K. R. Scherer & P. Ekman (Eds.), *Handbook of Method s in Nonverbal Behaviour Research* (pp. 136-198).

Scherer, K. R. (1984). On the nature and function of emotion: a component process approach. In K.R.Scherer & P.Ekman (Eds.), *Approaches to Emotion* (pp. 298-318). Hillsdale, NJ: Erlbaum.

Scherer, K. R. (1986). Vocal affect expression: a review and a model for future research. *Psychological Bulletin, 99*(2), 143-165.

Scherer, K. R. (1988a). Criteria for emotion-antecedent appraisal: a review. In V. Hamilton, G. H. Bower & N. H. Frijda (Eds.), *Cognitive Perspectives on Emotion and Motivation* (pp. 89-126). Dordrecht: Nijhoff.

Scherer, K. R. (1988b). On the symbolic functions of vocal affect expression. *Journal of Language and Social Psychology, 7*(2), 79-100.

Scherer, K. R. (1989). Vocal correlates of emotional arousal and affective disturbance. In H. Wagner & A. Manstead (Eds.), *Handbook of Social Psychophysiology* (pp. 165-197). New York: John Wiley & Sons Ltd.

Scherer, K. R. (1992). Vocal affect expression as symptom, symbol, and appeal. In H. Papousek, U. Jürgens & M. Papousek (Eds.), *Nonverbal Vocal Communication: Comparative and Developmental Approaches* (pp. 43-60). Cambridge: Cambridge University Press.

Scherer, K. R. (1994). Toward a concept of 'modal emotions'. In P. Ekman & R. J. Davidson (Eds.), *The Nature of Emotion: Fundamental Questions* (pp. 25-31). Oxford: Oxford University Press.

Scherer, K. R. (1999). Appraisal theory. In T. Dalgleish & M. Power (Eds.), *Handbook of Cognition and Emotion* (pp. 637-663). New York: John Wiley & Sons Ltd.

Scherer, K. R. (2000a). *A cross-cultural investigation of emotion inferences from voice and speech: implications for speech technology.* Paper presented at the 6th International Conference on Spoken Language Processing, Beijing, China.

Scherer, K. R. (2000b). Emotion. In M. Hewstone & W. Stroebe (Eds.), *Introduction to Social Psychology: a European Perspective* (3rd ed., pp. 151-191). Oxford: Blackwell.

Scherer, K. R. (2000c). Psychological models of emotion. In J. C. Borod (Ed.), *The Neuropsychology of Emotion* (pp. 137-162). Oxford: Oxford University Press.

Scherer, K. R. (2003). Vocal communication of emotion: a review of research paradigms. *Speech Communication, 40*, 227-256.

Scherer, K. R. (2004). Ways to study the nature and frequency of our daily emotions: reply to the commentaries on 'Emotions in everyday life'. *Social Science Information, 43*(3), 667-689.

Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology, 32*(1), 76-92.

Scherer, K. R., Banse, R., Wallbott, H. G., & Goldbeck, T. (1991). Vocal cues in emotion encoding and decoding. *Motivation and Emotion, 15*(2), 123-148.

Scherer, K. R., & Brosch, T. (2009). Culture-specific appraisal biases contribute to emotion dispositions. *European Journal of Personality, 23*, 265-288.

Scherer, K. R., Johnstone, T., & Klasmeyer, G. (2003). Vocal expression of emotion. In R. J. Davidson, K. R. Scherer & H. H. Goldsmith (Eds.), *Handbook of Affective Sciences* (pp. 433-456). Oxford: Oxford University Press.

Scherer, K. R., Ladd, D. R., & Silverman, K. E. A. (1984). Vocal cues to speaker affect: testing two models. *The Journal of the Acoustical Society of America, 76*(5), 1346-1356.

Scherer, K. R., & Wallbott, H. G. (1994). Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology, 66*(2), 310-328.

Scherer, K. R., Wranik, T., Sangsue, J., Tran, V., & Scherer, U. (2004). Emotions in everyday life: probability of occurrence, risk factors, appraisal and reaction patterns. *Social Science Information, 43*(4), 499-570.

Schlosberg, H. (1954). Three dimensions of emotion. *Psychological Review, 61*(2), 81-88.

Schröder, M. (1999). Can emotions be synthesized without controlling voice quality? *Phonus 4. Research Report of the Institute of Phonetics, Saarland University*, 37-55.

Schröder, M. (2001). *Emotional speech synthesis.* Paper presented at the Eurospeech 2001, Aalborg, Denmark.

Schröder, M. (2004a). Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions. In E. André, L. Dybkjaer, W. Minker & P. Heisterkamp (Eds.), *Affective Dialogue Systems* (Vol. 3068, pp. 209-220). Kloster Irsee, Germany.

Schröder, M. (2004b). Speech and Emotion Research: an Overview of Frameworks and a Dimensional Approach to Emotional Speech Synthesis. PhD thesis. *Phonus 7. Research Report of the Institute of Phonetics, Saarland University.*

Schröder, M. (2009). Approaches to emotional expressivity in synthetic speech. In K. Izdebski (Ed.), *Emotions in the Human Voice* (Vol. 3, pp. 307-321). San Diego, CA: Plural Publishing.

Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., & Gielen, S. (2001). *Acoustic correlates of emotion dimensions in view of speech synthesis.* Paper presented at the Eurospeech 2001, Aalborg.

Schröder, M., Heylen, D., & Poggi, I. (2006). *Perception of non-verbal emotional listener feedback.* Paper presented at the Speech Prosody 2006, Dresden, Germany.

Seki, K., Matsumoto, D., & Imahori, T. T. (2002). The conceptualisation and expression of intimacy in Japan and the United States. *Journal of Cross-Cultural Psychology, 33*(3), 303-319.

Seppänen, T., Toivanen, J., & Väyrynen, E. (2003). *MediaTeam speech corpus: a first large Finnish emotional database.* Paper presented at the XVth International Congress of Phonetic Sciences, Barcelona, Spain.

Shahid, S., Krahmer, E., & Swerts, M. (2008). *Real vs. acted emotional speech: comparing South-Asian and Caucasian speakers and observers.* Paper presented at the Speech Prosody 2008, Campinas, Brazil.

Shaver, P. R., Murdaya, U., & Fraley, R. C. (2001). Structure of the Indonesian Emotion Lexicon. *Asian Journal of Social Psychology, 4*(3), 201-224.

Shaver, P. R., Schwartz, J., Kirson, D., & O'Connor, C. (1987). Emotion knowledge: further exploration of a prototype approach. *Journal of Personality and Social Psychology, 52*(6), 1061-1086.

Shochi, T., Aubergé, V., & Rilliard, A. (2007). *Cross-listening of Japanese, English and French social affect: about universals, false friends and unknown attitudes.* Paper presented at the XVI International Congress of Phonetic Sciences, Saarbrücken, Germany.

Shrivastav, R. (2003). The use of an auditory model in predicting perceptual ratings of breathy voice quality. *Journal of Voice, 17*(4), 502-512.

Shrivastav, R. (2006). Multidimensional scaling of breathy voice quality: individual differences in perception. *Journal of Voice, 20*(2), 211-222.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlation: uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428.

Siegman, A. W., & Boyle, S. (1993). Voices of fear and anxiety and sadness and depression: the effects of speech rate and loudness on fear and anxiety and sadness and depression. *Journal of Abnormal Psychology, 102*(3), 430-437.

Skovenborg, E., & Nielsen, S. H. (2004). *Evaluation of different loudness models with music and speech material.* Paper presented at the 117th Convention of Audio Engineering Society, San Francisco, CA, USA.

Skovenborg, E., Quesnel, R., & Nielsen, S. H. (2004). *Loudness assessment of music and speech.* Paper presented at the 116th Convention of Audio Engineering Society, Berlin, Germany.

Sluijter, A. M. C., Shattuck-Hufnagel, S., Stevens, K. N., & Van Heuven, V. J. (1995). *Supralaryngeal resonance and glottal pulse shape as correlate of stress and accent in English.* Paper presented at the XIIIth International Congress of Phonetic Sciences, Stockholm, Sweden.

Sluijter, A. M. C., & van Heuven, V. J. (1996). Spectral balance as an acoustic correlate of linguistic stress. *The Journal of the Acoustical Society of America, 100*(4), 2471-2485.

Sluijter, A. M. C., van Heuven, V. J., & Pacilly, J. J. (1997). Spectral balance as a cue in the perception of linguistic stress. *The Journal of the Acoustical Society of America, 101*(1), 503-513.

Sobin, C., & Alpert, M. (1999). Emotion in speech: the acoustic attributes of fear, anger, sadness and joy. *Journal of Psycholinguistic Research, 28*(4), 347-365.

Sondhi, M. M. (1975). Measurement of the glottal waveform. *The Journal of the Acoustical Society of America, 57*(1), 228-232.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review, 64*, 153-181.

Streiner, D. L., & Norman, G. R. (2003). *Health Measurement Scales* (3 ed.). Oxford: Oxford University Press.

Streiner, D. L., & Norman, G. R. (2008). *Health Measurement Scales* (4 ed.). Oxford: Oxford University Press.

Svetozarova, N. (1998). Intonation in Russian. In D. Hirst & A. Di Cristo (Eds.), *Intonation Systems* (pp. 261-274). Cambridge: Cambridge University Press.

Swerts, M., & Veldhuis, R. (2001). The effect of speech melody on voice quality. *Speech Communication, 33*(4), 297-303.

Taylor, P. (2009). *Text-to-Speech Synthesis*. Cambridge: Cambridge University Press.

Teshigawara, M. (2004). *Random splicing: a method of investigating the effects of voice quality on impression formation.* Paper presented at the Speech Prosody 2004, Nara, Japan.

Teshigawara, M., Amir, N., Amir, O., Milano Wlosko, E., & Avivi, M. (2009). Perception of Japanese *anime* voices by Hebrew speakers. In K. Izdebski (Ed.), *Emotions in the Human Voice* (Vol. 3, pp. 177-186). San Diego, CA: Plural Publishing.

't Hart, J., Collier, R., & Cohen, A. (1990). *A Perceptual Study of Intonation: an Experimental-Phonetic Approach to Speech Melody.* Cambridge: Cambridge University Press.

Titze, I. R. (1994). *Principles of Voice Production.* Englewood Cliffs, New Jersey: Prentice Hall.

Toivanen, J., Seppänen, T., & Väyrynen, E. (2009). Automatic discrimination of emotion from voice: a review of research paradigms. In K. Izdebski (Ed.), *Emotions in the Human Voice* (Vol. 3, pp. 67-77). San Diego, CA: Plural Publishing.

Toivanen, J., Waaramaa, T., Alku, P., Laukkanen, A.-M., Seppänen, T., Väyrynen, E., et al. (2006). Emotions in [a]: a perceptual and acoustic study. *Logopedics Phoniatrics Vocology, 31*, 43-48.

Tolkmitt, F. J., & Scherer, K. R. (1986). Effects of experimentally induced stress on vocal parameters. *Journal of Experimental Psychology: Human Perception and Performance, 12*(3), 302-313.

Tomkins, S. S. (1995). *Exploring Affect: the Selected Writings of Sylvan S. Tomkins.* Cambridge: Cambridge University Press.

Tooher, M., Yanushevskaya, I., & Gobl, C. (2008). *Transformation of LF parameters for speech synthesis of emotion: regression trees.* Paper presented at the Speech Prosody 2008, Campinas, Brazil.

Traunmüller, H., & Eriksson, A. (2000). Acoustic effects of variation in vocal effort by men, women, and children. *The Journal of the Acoustical Society of America, 107*(6), 3438-3451.

Trittin, P. J., & de Santos y Lleó, A. (1995). Voice quality analysis of male and female Spanish speakers. *Speech Communication, 16*, 359-368.

Turk, A. E., & Sawusch, J. R. (1996). The processing of duration and intensity cues to prominence. *The Journal of the Acoustical Society of America, 99*(6), 3782-3790.

Uldall, E. (1964). Dimensions of meaning in intonation. In D. Abercrombie, D. B. Fry, P. A. D. MacCarthy, M. C. Scott & J. L. M. Trim (Eds.), *In Honour of Daniel Jones:*

papers contributed on the occasion of his eightieth birthday, 12 September 1961 (pp. 271-279).

van Bezooijen, R., Otto, S. A., & Heenan, T. A. (1983). Recognition of vocal expressions of emotion: a three-nation study to identify universal characteristics. *Journal of Cross-Cultural Psychology, 14*(4), 387-406.

Velten, E. (1968). A laboratory task for induction of mood states. *Behaviour Research and Therapy, 6*, 473-482.

Verhey, J. L., & Kollmeier, B. (2002). Spectral loudness summation as a function of duration. *The Journal of the Acoustical Society of America, 111*(3), 1349-1358.

Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication, 48*(9), 1162-1181.

Vinden, P. G., & Astington, J. W. (2000). Culture and understanding other minds. In S. Baron-Cohen, H. Tager-Flusberg & D. J. Cohen (Eds.), *Understanding Other Minds* (2 ed., pp. 503-519). Oxford: Oxford University Press.

Vysotskaya, L. G., & Kovyazina, M. A. (2002). The comparative analysis of the intonation of basic positive emotions in English and Russian spontaneous speech. *Language and Literature, 17*.

Waaramaa, T., Alku, P., & Laukkanen, A.-M. (2006). The role of F3 in the vocal expression of emotion. *Logopedics Phoniatrics Vocology, 31*, 153-156.

Waaramaa, T., Laukkanen, A., Alku, P., & Väyrynen, E. (2008). Monopitched expression of emotions in different vowels. *Folia Phoniatrica et Logopaedica, 60*(5), 249-255.

Wagner, H. L. (1993). On measuring performance in category judgment studies of nonverbal behavior. *Journal of Nonverbal Behaviour, 17*, 2-28.

Walbott, H. G., Ricci-Bitti, P., & Bänninger-Huber, E. (1986). Non-verbal reactions to emotional experiences. In K.R.Scherer, H.G.Walbott & A.B.Summerfield (Eds.), *Experiencing Emotion: a Cross-Cultural Study* (pp. 98-116). Cambridge: Cambridge University Press.

Wells, J. C. (1982). *Accents of English*. Cambridge: Cambridge University Press.

Wells, J. C. (2006). *English Intonation: an Introduction*. Cambridge: Cambridge University Press.

Westermann, R., Spies, K., Stahl, G., & Hesse, F. W. (1996). Relative effectiveness and validity of mood induction procedures: a meta-analysis. *European Journal of Social Psychology, 26*, 557-580.

Wichmann, A. (2000). *The attitudinal effects of prosody, and how they relate to emotion.* Paper presented at the ITRW on Speech and Emotion, Newcastle, Northern Ireland.

Williams, C. E., & Stevens, K. N. (1972). Emotions and speech: some acoustical correlates. *The Journal of the Acoustical Society of America, 52*(4), 1238-1250.

Wilting, J., Krahmer, E., & Swerts, M. (2006). *Real vs. acted emotional speech.* Paper presented at the Interspeech 2006, Pittsburgh, Pennsylvania, USA.

Yaffe, R. A. (1998). Enhancement of reliability analysis: application of intraclass correlations with SPSS/Windows v.8 http://www.nyu.edu/its/statistics/Docs/intracls/html.

Yanushevskaya, I., Gobl, C., & Ní Chasaide, A. (2005). *Voice quality and f0 cues for affect expression: implications for synthesis.* Paper presented at the Interspeech 2005 - Eurospeech, Lisbon, Portugal.

Yanushevskaya, I., Tooher, M., Gobl, C., & Ní Chasaide, A. (2007). Time- and amplitude-based voice source correlates of emotional portrayals. In A. Paiva, R. Prada & R. W. Picard (Eds.), *Affective Computing and Intelligent Interaction: Proceedings of the ACII 2007* (Vol. 4738, pp. 159-170). Lisbon, Portugal: Springer-Verlag.

Yuan, J., Shen, L., & Chen, F. (2002). *The acoustic realisation of anger, fear, joy and sadness in Chinese.* Paper presented at the 7th International Conference on Spoken Language Processing, Denver, Colorado, USA.

Zara, A., Maffiolo, V., Martin, J.-C., & Devillers, L. (2007). Collection and annotation of a corpus of human-human multimodal interactions: emotion and other anthropomorphic characteristics. In A. Paiva, R. Prada & R. W. Picard (Eds.), *Affective Computing and Intelligent Interaction: Proceedings of the ACII 2007* (Vol. 4738, pp. 464-475). Lisbon: Springer-Verlag.

Zinken, J., Knoll, M., & Panksepp, J. (2008). Universality and diversity in the vocalisation of emotions. In K. Izdebski (Ed.), *Emotions in the Human Voice* (pp. 185-202). San Diego, CA: Plural Publishing.

Zraick, R. I., Marshall, W., Smith-Olinde, L., & Montague, J. C. (2004). The effect of task on determination of habitual loudness. *Journal of Voice, 18*(2), 176-182.

Zwicker, E., Flottorp, G., & Stevens, S. S. (1957). Critical bandwidth in loudness summation. *The Journal of the Acoustical Society of America, 29*, 548-557.

1 Appendix 1

Ratings obtained in the cross-language study (Chapter 4).

For each stimulus, in the box plots shown are mean (dot), median (line), interquartile range (box) and extreme values (whiskers).

Red rectangles highlight the cases where affective ratings yielded by the stimuli were above 1, corresponding to mild to strong affective colouring.

Figures 1.1-1.2 show results for Hiberno-English listeners.

Figures 1.3-1.4 show results for Russian listeners.

Figures 1.5-1.6 show results for Spanish listeners.

Figures 1.7-1.8 show results for Japanese listeners.

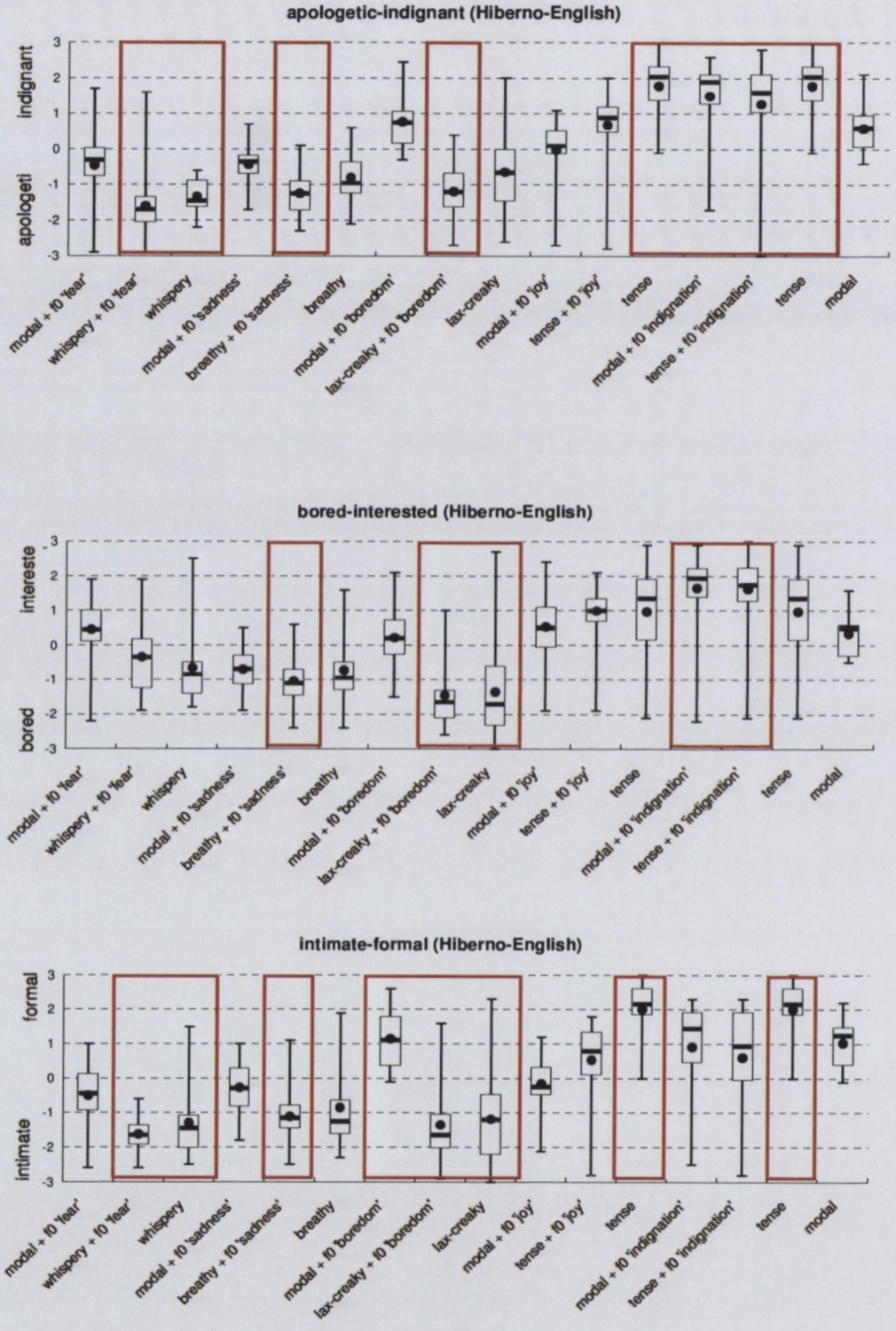


Figure 1.1. *Mean rating obtained for the stimuli for each sub-test (Hiberno-English listeners). The plots show interquartile range (box), mean (filled circle), median (horizontal line in box) and extreme values (whiskers).*

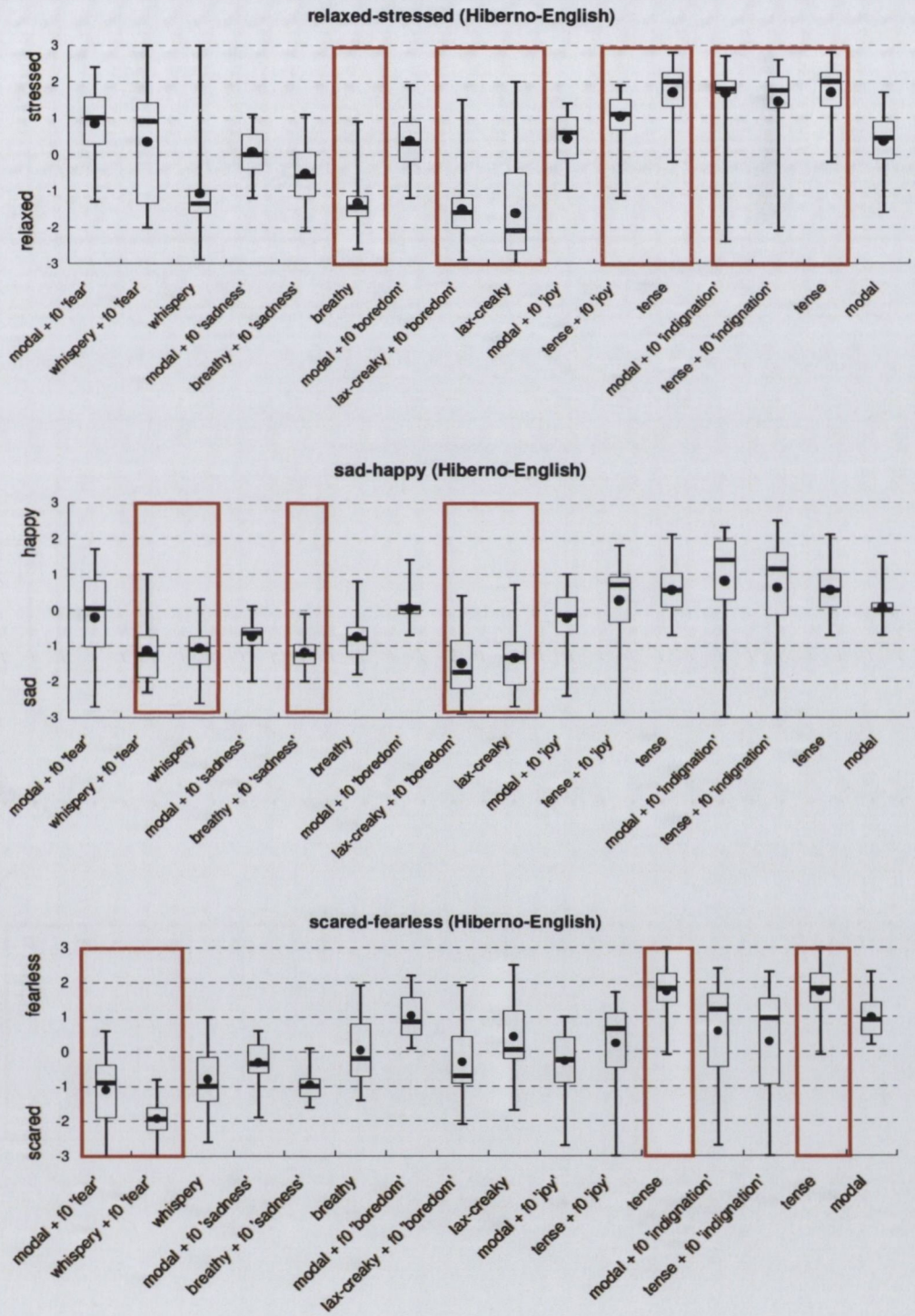


Figure 1.2. *Mean ratings obtained for the stimuli for each sub-test (Hiberno-English listeners) (continued). The plots show interquartile range (box), mean (filled circle), median (horizontal line in box) and extreme values (whiskers).*

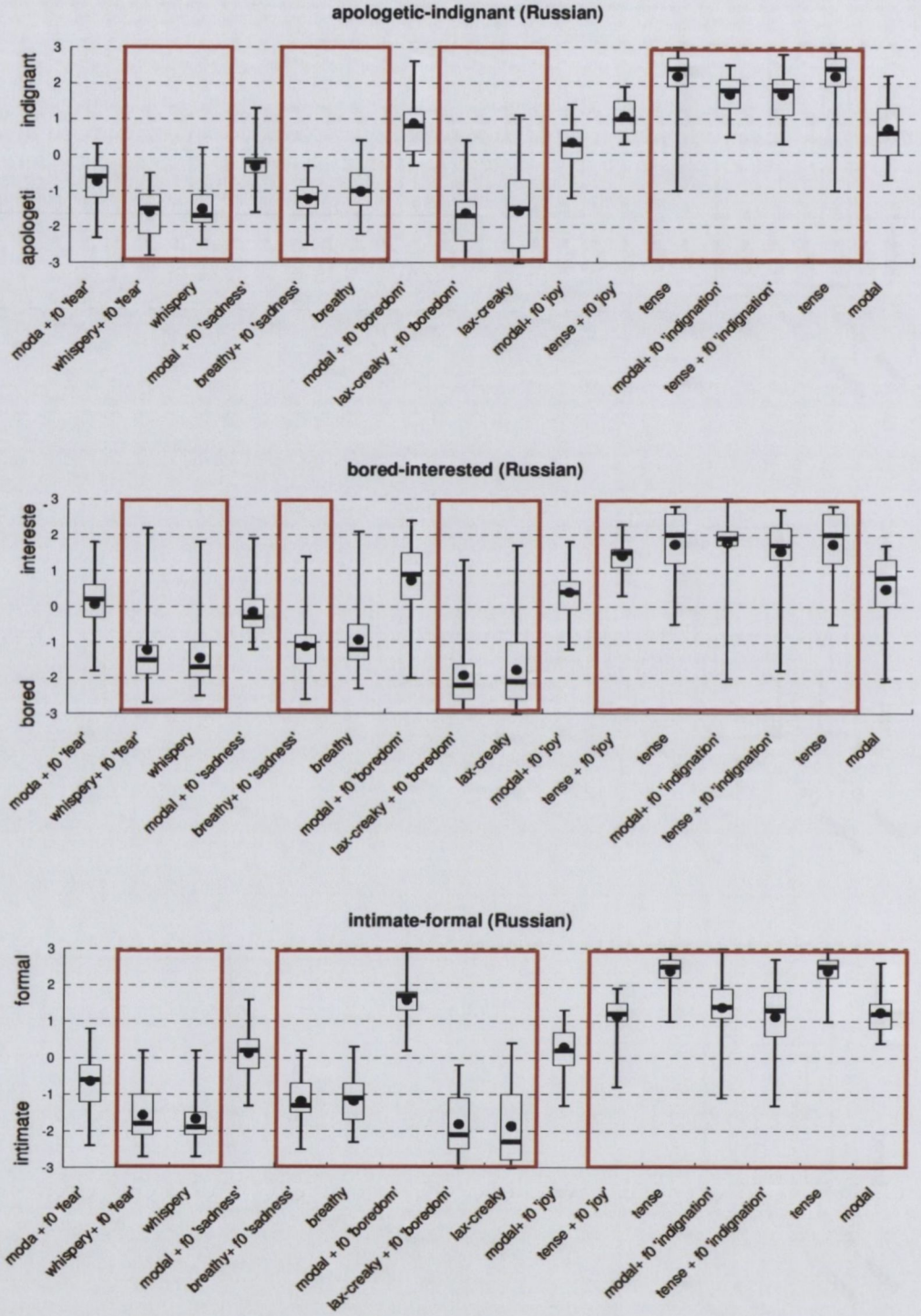


Figure 1.3. *Mean rating obtained for the stimuli for each sub-test (Russian listeners). The plots show interquartile range (box), mean (filled circle), median (horizontal line in box) and extreme values (whiskers).*

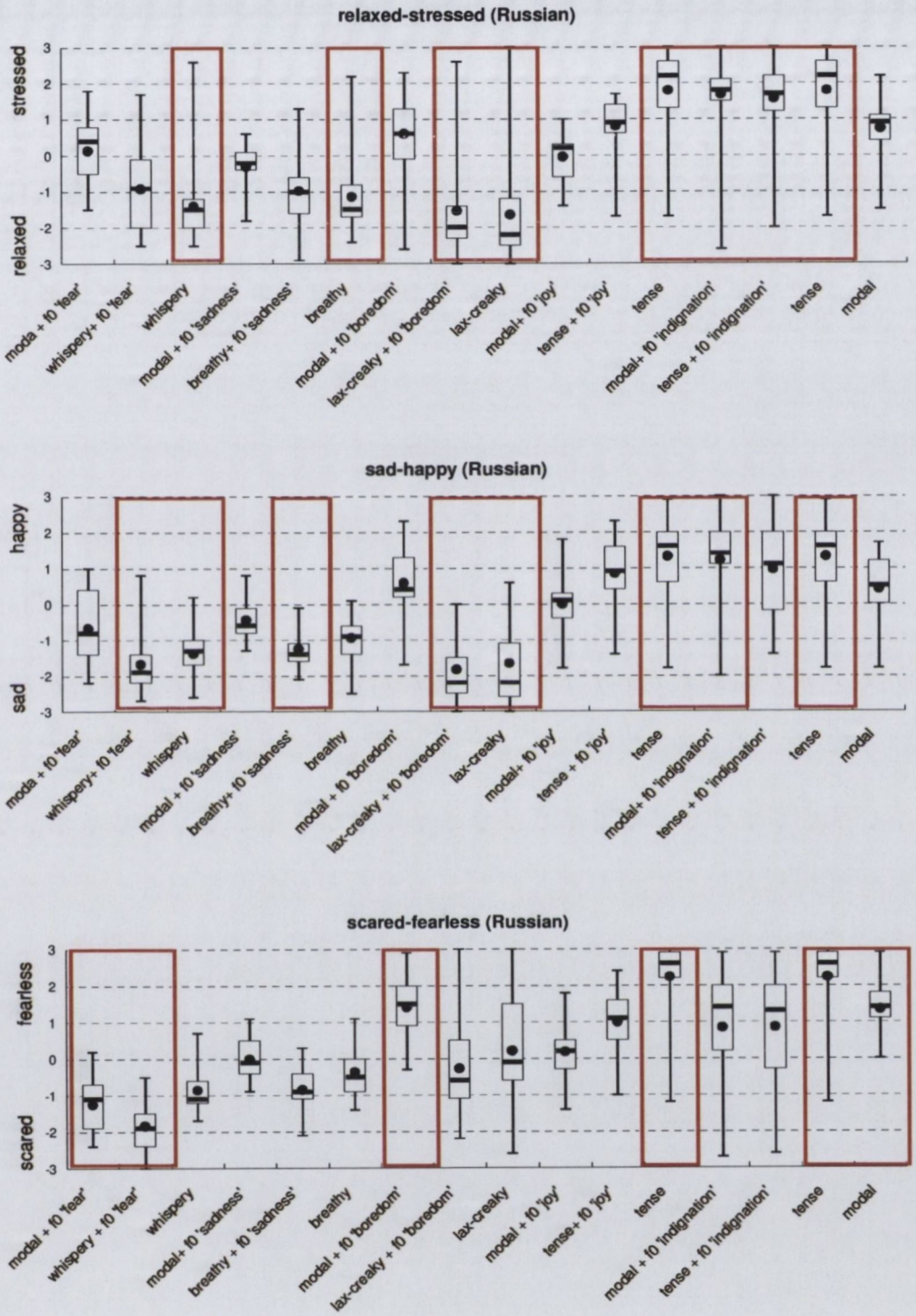


Figure 1.4. *Mean ratings obtained for the stimuli for each sub-test (Russian listeners) (continued). The plots show interquartile range (box), mean (filled circle), median (horizontal line in box) and extreme values (whiskers).*

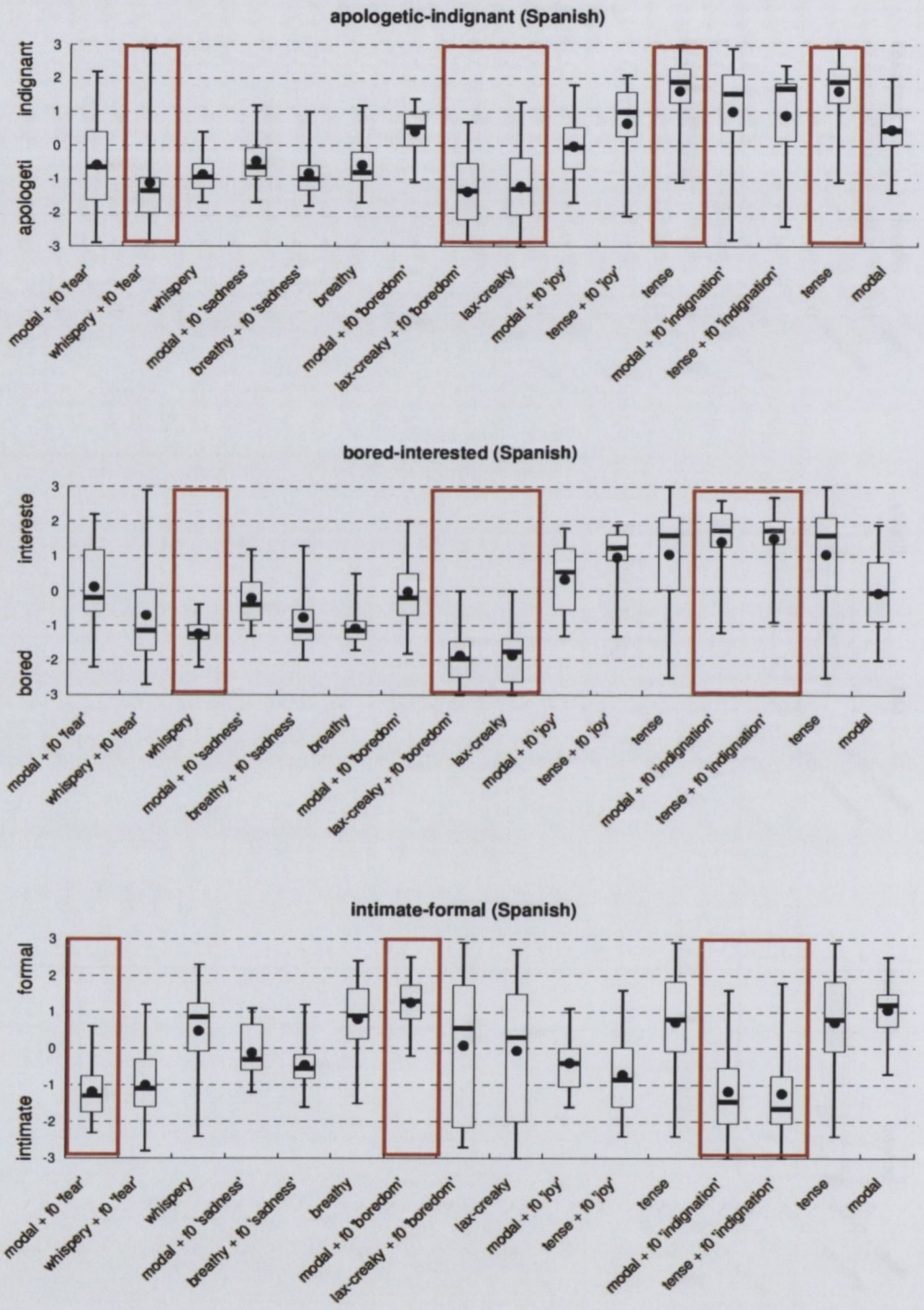


Figure 1.5. *Mean rating obtained for the stimuli for each sub-test (Spanish listeners). The plots show interquartile range (box), mean (filled circle), median (horizontal line in box) and extreme values (whiskers).*

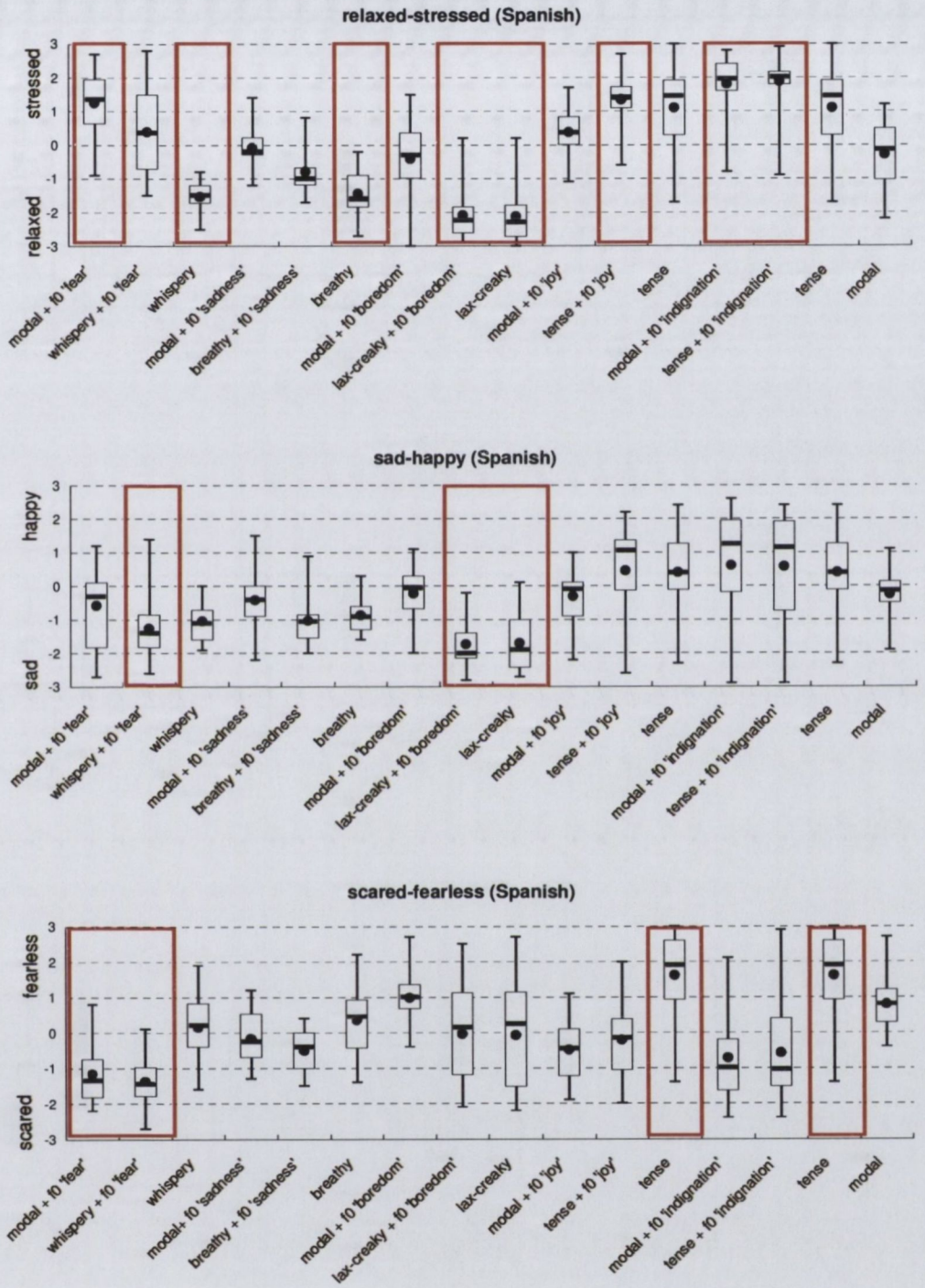


Figure 1.6. *Mean ratings obtained for the stimuli for each sub-test (Spanish listeners) (continued). The plots show interquartile range (box), mean (filled circle), median (horizontal line in box) and extreme values (whiskers).*

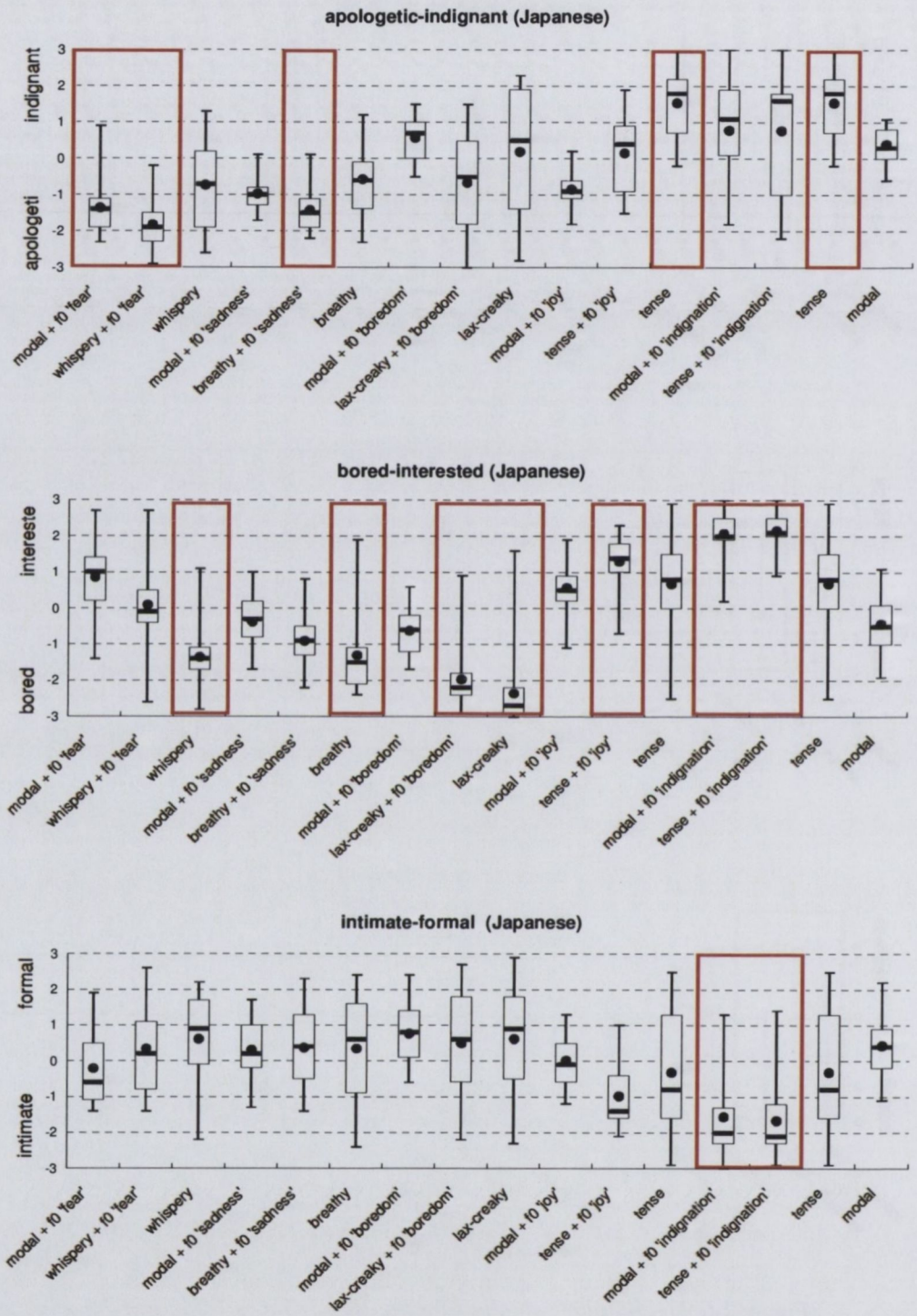


Figure 1.7. *Mean rating obtained for the stimuli for each sub-test (Spanish listeners). The plots show interquartile range (box), mean (filled circle), median (horizontal line in box) and extreme values (whiskers).*

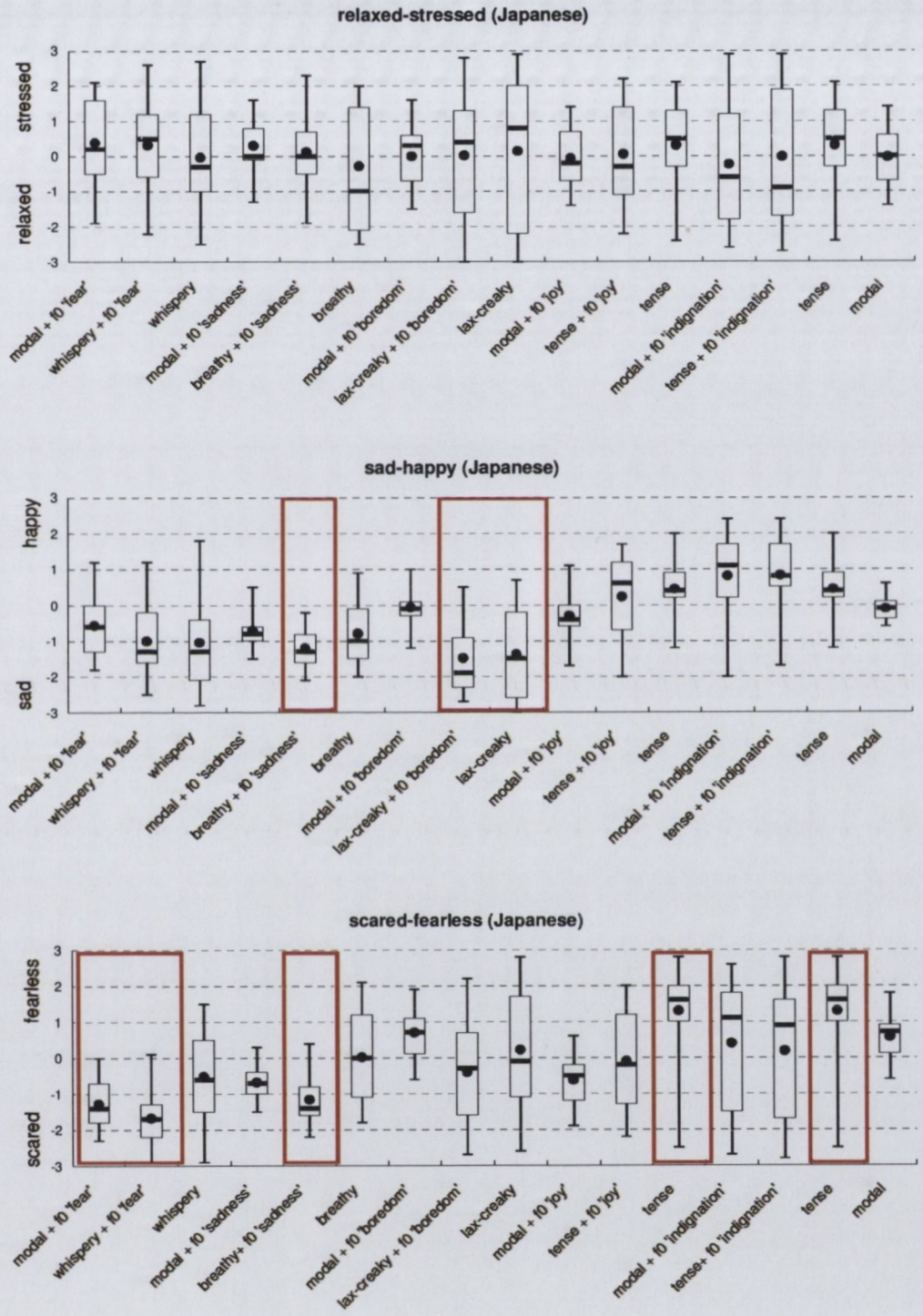


Figure 1.8. *Mean ratings obtained for the stimuli for each sub-test (Japanese listeners) (continued). The plots show interquartile range (box), mean (filled circle), median (horizontal line in box) and extreme values (whiskers).*

2 Appendix 2

Results of the omnibus 'group' ANOVA for the cross-language study (Chapter 4).

Apologetic-Indignant Part 1

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Target emotion group	Sphericity Assumed	514.424	3	171.475	123.669	.000	.613
	Greenhouse-Geisser	514.424	1.998	257.519	123.669	.000	.613
Target emotion group * Language	Sphericity Assumed	62.149	9	6.905	4.980	.000	.161
	Greenhouse-Geisser	62.149	5.993	10.370	4.980	.000	.161
Error(Target emotion group)	Sphericity Assumed	324.455	234	1.387			
	Greenhouse-Geisser	324.455	155.814	2.082			
Stimulus type	Sphericity Assumed	81.492	2	40.746	35.477	.000	.313
	Greenhouse-Geisser	81.492	1.343	60.700	35.477	.000	.313
Stimulus type* language	Sphericity Assumed	39.266	6	6.544	5.698	.000	.180
	Greenhouse-Geisser	39.266	4.028	9.749	5.698	.000	.180
Error(Stimulus type)	Sphericity Assumed	179.169	156	1.149			
	Greenhouse-Geisser	179.169	104.719	1.711			
Target emotion group* Stimulus type	Sphericity Assumed	270.174	6	45.029	110.736	.000	.587
	Greenhouse-Geisser	270.174	3.593	75.195	110.736	.000	.587
Target emotion group* Stimulus type* Language	Sphericity Assumed	11.387	18	.633	1.556	.067	.056
	Greenhouse-Geisser	11.387	10.779	1.056	1.556	.113	.056
Error(Target emotion group* Stimulus type)	Sphericity Assumed	190.305	468	.407			
	Greenhouse-Geisser	190.305	280.253	.679			

Tests of Between-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	169.887	1	169.887	132.577	.000	.630
Language	2.755	3	.918	.717	.545	.027
Error	99.951	78	1.281			

Apologetic-Indignant Part 2

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Target emotion group * Language	Sphericity Assumed	954.001	3	318.000	164.920	.000	.679
	Greenhouse-Geisser	954.001	1.896	503.079	164.920	.000	.679
Error(Target emotion group)	Sphericity Assumed	61.334	9	6.815	3.534	.000	.120
	Greenhouse-Geisser	61.334	5.689	10.781	3.534	.003	.120
Stimulus type	Sphericity Assumed	451.202	234	1.928			
	Greenhouse-Geisser	451.202	147.913	3.050			
Stimulus type* language	Sphericity Assumed	115.166	2	57.583	41.689	.000	.348
	Greenhouse-Geisser	115.166	1.393	82.696	41.689	.000	.348
Error(Stimulus type)	Sphericity Assumed	34.802	6	5.800	4.199	.001	.139
	Greenhouse-Geisser	34.802	4.178	8.330	4.199	.003	.139
Target emotion group* Stimulus type	Sphericity Assumed	215.473	156	1.381			
	Greenhouse-Geisser	215.473	108.626	1.984			
Target emotion group* Stimulus type* Language	Sphericity Assumed	102.308	6	17.051	43.650	.000	.359
	Greenhouse-Geisser	102.308	3.551	28.809	43.650	.000	.359
Error(Target emotion group* Stimulus type)	Sphericity Assumed	14.104	18	.784	2.006	.008	.072
	Greenhouse-Geisser	14.104	10.654	1.324	2.006	.030	.072
Target emotion group * Language	Sphericity Assumed	182.820	468	.391			
	Greenhouse-Geisser	182.820	277.000	.660			

Tests of Between-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	65.418	1	65.418	59.806	.000	.434
Language	1.534	3	.511	.468	.706	.018
Error	85.318	78	1.094			

Bored-Interested Part 1

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Target emotion group * Language	Sphericity Assumed	610.944	3	203.648	174.467	.000	.691
	Greenhouse-Geisser	610.944	2.285	267.316	174.467	.000	.691
Error(Target emotion group)	Sphericity Assumed	46.629	9	5.181	4.439	.000	.146
	Greenhouse-Geisser	46.629	6.856	6.801	4.439	.000	.146
Stimulus type	Sphericity Assumed	273.138	234	1.167			
	Greenhouse-Geisser	273.138	178.267	1.532			
Stimulus type* language	Sphericity Assumed	134.971	2	67.485	62.837	.000	.446
	Greenhouse-Geisser	134.971	1.587	85.040	62.837	.000	.446
Error(Stimulus type)	Sphericity Assumed	20.688	6	3.448	3.210	.005	.110
	Greenhouse-Geisser	20.688	4.761	4.345	3.210	.010	.110
Target emotion group* Stimulus type	Sphericity Assumed	167.541	156	1.074			
	Greenhouse-Geisser	167.541	123.798	1.353			
Target emotion group* Stimulus type* Language	Sphericity Assumed	208.512	6	34.752	63.354	.000	.448
	Greenhouse-Geisser	208.512	2.448	85.161	63.354	.000	.448
Error(Target emotion group* Stimulus type)	Sphericity Assumed	20.873	18	1.160	2.114	.005	.075
	Greenhouse-Geisser	20.873	7.345	2.842	2.114	.041	.075
Target emotion group * Language	Sphericity Assumed	256.713	468	.549			
	Greenhouse-Geisser	256.713	190.978	1.344			

Tests of Between-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	138.370	1	138.370	56.312	.000	.419
Language	5.661	3	1.887	.768	.515	.029
Error	191.661	78	2.457			

Bored-Interested Part 2

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Target emotion group * Language	Sphericity Assumed	1061.461	3	353.820	221.333	.000	.739
	Greenhouse-Geisser	1061.461	2.004	529.579	221.333	.000	.739
Error(Target emotion group)	Sphericity Assumed	46.987	9	5.221	3.266	.001	.112
	Greenhouse-Geisser	46.987	6.013	7.814	3.266	.005	.112
Stimulus type	Sphericity Assumed	374.070	234	1.599			
	Greenhouse-Geisser	374.070	156.339	2.393			
Stimulus type* language	Sphericity Assumed	245.174	2	122.587	101.733	.000	.566
	Greenhouse-Geisser	245.174	1.722	142.390	101.733	.000	.566
Error(Stimulus type)	Sphericity Assumed	30.325	6	5.054	4.194	.001	.139
	Greenhouse-Geisser	30.325	5.166	5.871	4.194	.001	.139
Target emotion group* Stimulus type	Sphericity Assumed	187.979	156	1.205			
	Greenhouse-Geisser	187.979	134.305	1.400			
Target emotion group* Stimulus type* Language	Sphericity Assumed	93.203	6	15.534	33.593	.000	.301
	Greenhouse-Geisser	93.203	2.964	31.448	33.593	.000	.301
Error(Target emotion group* Stimulus type)	Sphericity Assumed	18.553	18	1.031	2.229	.003	.079
	Greenhouse-Geisser	18.553	8.891	2.087	2.229	.022	.079
Target emotion group * Language	Sphericity Assumed	216.406	468	.462			
	Greenhouse-Geisser	216.406	231.171	.936			

Tests of Between-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	49.269	1	49.269	27.139	.000	.258
Language	4.668	3	1.556	.857	.467	.032
Error	141.601	78	1.815			

Intimate-Formal Part 1

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Target emotion group * Language	Sphericity Assumed	145.822	3	48.607	27.354	.000	.260
	Greenhouse-Geisser	145.822	1.945	74.965	27.354	.000	.260
Error(Target emotion group)	Sphericity Assumed	287.430	9	31.937	17.972	.000	.409
	Greenhouse-Geisser	287.430	5.836	49.255	17.972	.000	.409
Stimulus type	Sphericity Assumed	415.815	234	1.777			
	Greenhouse-Geisser	415.815	151.725	2.741			
Stimulus type* language	Sphericity Assumed	82.875	2	41.437	31.343	.000	.287
	Greenhouse-Geisser	82.875	1.424	58.197	31.343	.000	.287
Error(Stimulus type)	Sphericity Assumed	63.148	6	10.525	7.961	.000	.234
	Greenhouse-Geisser	63.148	4.272	14.782	7.961	.000	.234
Target emotion group* Stimulus type	Sphericity Assumed	206.244	156	1.322			
	Greenhouse-Geisser	206.244	111.074	1.857			
Target emotion group* Stimulus type* Language	Sphericity Assumed	207.047	6	34.508	56.092	.000	.418
	Greenhouse-Geisser	207.047	3.008	68.833	56.092	.000	.418
Error(Target emotion group* Stimulus type)	Sphericity Assumed	156.622	18	8.701	14.144	.000	.352
	Greenhouse-Geisser	156.622	9.024	17.356	14.144	.000	.352
Target emotion group * Language	Sphericity Assumed	287.914	468	.615			
	Greenhouse-Geisser	287.914	234.620	1.227			

Tests of Between-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	22.827	1	22.827	12.267	.001	.136
Language	58.093	3	19.364	10.406	.000	.286
Error	145.152	78	1.861			

Intimate-Formal Part 2

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Target emotion group * Language	Sphericity Assumed	119.366	3	39.789	16.529	.000	.175
	Greenhouse-Geisser	119.366	1.705	70.021	16.529	.000	.175
Error(Target emotion group)	Sphericity Assumed	553.708	9	61.523	25.558	.000	.496
	Greenhouse-Geisser	553.708	5.114	108.270	25.558	.000	.496
Stimulus type	Sphericity Assumed	563.293	234	2.407			
	Greenhouse-Geisser	563.293	132.968	4.236			
Stimulus type* language	Sphericity Assumed	100.672	2	50.336	32.014	.000	.291
	Greenhouse-Geisser	100.672	1.416	71.112	32.014	.000	.291
Error(Stimulus type)	Sphericity Assumed	126.497	6	21.083	13.409	.000	.340
	Greenhouse-Geisser	126.497	4.247	29.785	13.409	.000	.340
Target emotion group* Stimulus type	Sphericity Assumed	245.276	156	1.572			
	Greenhouse-Geisser	245.276	110.423	2.221			
Target emotion group* Stimulus type* Language	Sphericity Assumed	214.322	6	35.720	62.964	.000	.447
	Greenhouse-Geisser	214.322	2.958	72.446	62.964	.000	.447
Error(Target emotion group* Stimulus type)	Sphericity Assumed	51.930	18	2.885	5.085	.000	.164
	Greenhouse-Geisser	51.930	8.875	5.851	5.085	.000	.164
Target emotion group * Language	Sphericity Assumed	265.504	468	.567			
	Greenhouse-Geisser	265.504	230.753	1.151			

Tests of Between-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	31.898	1	31.898	22.833	.000	.226
Language	14.554	3	4.851	3.473	.020	.118
Error	108.968	78	1.397			

Relaxed-Stressed Part 1

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Target emotion group * Language	Sphericity Assumed	338.831	3	112.944	57.386	.000	.424
	Greenhouse-Geisser	338.831	2.100	161.321	57.386	.000	.424
Error(Target emotion group)	Sphericity Assumed	139.851	9	15.539	7.895	.000	.233
	Greenhouse-Geisser	139.851	6.301	22.195	7.895	.000	.233
Stimulus type	Sphericity Assumed	460.547	234	1.968			
	Greenhouse-Geisser	460.547	163.827	2.811			
Stimulus type* language	Sphericity Assumed	102.058	2	51.029	34.118	.000	.304
	Greenhouse-Geisser	102.058	1.549	65.873	34.118	.000	.304
Error(Stimulus type)	Sphericity Assumed	38.670	6	6.445	4.309	.000	.142
	Greenhouse-Geisser	38.670	4.648	8.320	4.309	.002	.142
Target emotion group* Stimulus type	Sphericity Assumed	233.323	156	1.496			
	Greenhouse-Geisser	233.323	120.846	1.931			
Target emotion group* Stimulus type* Language	Sphericity Assumed	219.207	6	36.535	57.992	.000	.426
	Greenhouse-Geisser	219.207	2.791	78.546	57.992	.000	.426
Error(Target emotion group* Stimulus type)	Sphericity Assumed	65.451	18	3.636	5.772	.000	.182
	Greenhouse-Geisser	65.451	8.372	7.817	5.772	.000	.182
Target emotion group * Language	Sphericity Assumed	294.839	468	.630			
	Greenhouse-Geisser	294.839	217.685	1.354			

Tests of Between-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	32.319	1	32.319	15.380	.000	.165
Language	38.716	3	12.905	6.142	.001	.191
Error	163.903	78	2.101			

Relaxed-Stressed Part 2

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Target emotion group * Language	Sphericity Assumed	602.457	3	200.819	68.642	.000	.468
	Greenhouse-Geisser	602.457	1.738	346.596	68.642	.000	.468
Error(Target emotion group)	Sphericity Assumed	251.058	9	27.895	9.535	.000	.268
	Greenhouse-Geisser	251.058	5.215	48.145	9.535	.000	.268
Stimulus type	Sphericity Assumed	684.590	234	2.926			
	Greenhouse-Geisser	684.590	135.581	5.049			
Stimulus type* language	Sphericity Assumed	182.548	2	91.274	48.721	.000	.384
	Greenhouse-Geisser	182.548	1.585	115.167	48.721	.000	.384
Error(Stimulus type)	Sphericity Assumed	64.781	6	10.797	5.763	.000	.181
	Greenhouse-Geisser	64.781	4.755	13.623	5.763	.000	.181
Target emotion group* Stimulus type	Sphericity Assumed	292.254	156	1.873			
	Greenhouse-Geisser	292.254	123.636	2.364			
Target emotion group* Stimulus type* Language	Sphericity Assumed	92.746	6	15.458	34.425	.000	.306
	Greenhouse-Geisser	92.746	3.519	26.354	34.425	.000	.306
Error(Target emotion group* Stimulus type)	Sphericity Assumed	33.785	18	1.877	4.180	.000	.139
	Greenhouse-Geisser	33.785	10.558	3.200	4.180	.000	.139
Target emotion group * Language	Sphericity Assumed	210.146	468	.449			
	Greenhouse-Geisser	210.146	274.502	.766			

Tests of Between-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	3.488	1	3.488	2.529	.116	.031
Language	14.179	3	4.726	3.426	.021	.116
Error	107.593	78	1.379			

Sad-Happy Part 1

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Target emotion group * Language	Sphericity Assumed	295.641	3	98.547	85.397	.000	.523
	Greenhouse-Geisser	295.641	1.931	153.113	85.397	.000	.523
Error(Target emotion group)	Sphericity Assumed	23.854	9	2.650	2.297	.017	.081
	Greenhouse-Geisser	23.854	5.793	4.118	2.297	.040	.081
Stimulus type	Sphericity Assumed	270.032	234	1.154			
	Greenhouse-Geisser	270.032	150.608	1.793			
Stimulus type* language	Sphericity Assumed	62.390	2	31.195	34.700	.000	.308
	Greenhouse-Geisser	62.390	1.429	43.673	34.700	.000	.308
Error(Stimulus type)	Sphericity Assumed	4.051	6	.675	.751	.609	.028
	Greenhouse-Geisser	4.051	4.286	.945	.751	.568	.028
Target emotion group* Stimulus type	Sphericity Assumed	140.243	156	.899			
	Greenhouse-Geisser	140.243	111.428	1.259			
Target emotion group* Stimulus type* Language	Sphericity Assumed	165.879	6	27.646	76.133	.000	.494
	Greenhouse-Geisser	165.879	3.056	54.287	76.133	.000	.494
Error(Target emotion group* Stimulus type)	Sphericity Assumed	12.621	18	.701	1.931	.012	.069
	Greenhouse-Geisser	12.621	9.167	1.377	1.931	.047	.069
Target emotion group * Language	Sphericity Assumed	169.947	468	.363			
	Greenhouse-Geisser	169.947	238.335	.713			

Tests of Between-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	397.368	1	397.368	192.110	.000	.711
Language	1.498	3	.499	.241	.867	.009
Error	161.338	78	2.068			

Sad-Happy Part 2

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Target emotion group * Language	Sphericity Assumed	546.237	3	182.079	107.236	.000	.579
	Greenhouse-Geisser	546.237	1.639	333.374	107.236	.000	.579
Error(Target emotion group)	Sphericity Assumed	22.906	9	2.545	1.499	.149	.055
	Greenhouse-Geisser	22.906	4.916	4.660	1.499	.196	.055
Stimulus type	Sphericity Assumed	397.316	234	1.698			
	Greenhouse-Geisser	397.316	127.804	3.109			
Stimulus type* language	Sphericity Assumed	120.385	2	60.193	50.202	.000	.392
	Greenhouse-Geisser	120.385	1.444	83.395	50.202	.000	.392
Error(Stimulus type)	Sphericity Assumed	6.677	6	1.113	.928	.477	.034
	Greenhouse-Geisser	6.677	4.331	1.542	.928	.456	.034
Target emotion group* Stimulus type	Sphericity Assumed	187.046	156	1.199			
	Greenhouse-Geisser	187.046	112.597	1.661			
Target emotion group* Stimulus type* Language	Sphericity Assumed	74.430	6	12.405	40.893	.000	.344
	Greenhouse-Geisser	74.430	3.797	19.604	40.893	.000	.344
Error(Target emotion group* Stimulus type)	Sphericity Assumed	9.833	18	.546	1.801	.023	.065
	Greenhouse-Geisser	9.833	11.390	.863	1.801	.051	.065
Target emotion group * Language	Sphericity Assumed	141.969	468	.303			
	Greenhouse-Geisser	141.969	296.136	.479			

Tests of Between-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	267.404	1	267.404	112.233	.000	.590
Language	2.299	3	.766	.322	.810	.012
Error	185.841	78	2.383			

Scared-Fearless Part 1

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Target emotion group * Language	Sphericity Assumed	427.110	3	142.370	90.753	.000	.538
	Greenhouse-Geisser	427.110	1.622	263.364	90.753	.000	.538
Error(Target emotion group)	Sphericity Assumed	39.021	9	4.336	2.764	.004	.096
	Greenhouse-Geisser	39.021	4.865	8.020	2.764	.022	.096
Stimulus type	Sphericity Assumed	367.093	234	1.569			
	Greenhouse-Geisser	367.093	126.496	2.902			
Stimulus type* language	Sphericity Assumed	164.958	2	82.479	77.170	.000	.497
	Greenhouse-Geisser	164.958	1.249	132.063	77.170	.000	.497
Error(Stimulus type)	Sphericity Assumed	7.644	6	1.274	1.192	.313	.044
	Greenhouse-Geisser	7.644	3.747	2.040	1.192	.319	.044
Target emotion group* Stimulus type	Sphericity Assumed	166.732	156	1.069			
	Greenhouse-Geisser	166.732	97.429	1.711			
Target emotion group* Stimulus type* Language	Sphericity Assumed	175.849	6	29.308	63.851	.000	.450
	Greenhouse-Geisser	175.849	3.051	57.635	63.851	.000	.450
Error(Target emotion group* Stimulus type)	Sphericity Assumed	14.999	18	.833	1.815	.021	.065
	Greenhouse-Geisser	14.999	9.153	1.639	1.815	.065	.065
Target emotion group * Language	Sphericity Assumed	214.816	468	.459			
	Greenhouse-Geisser	214.816	237.982	.903			

Tests of Between-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	25.467	1	25.467	13.443	.000	.147
Language	14.829	3	4.943	2.609	.057	.091
Error	147.764	78	1.894			

Scared-Fearless Part 2

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Target emotion group * Language	Sphericity Assumed	494.132	3	164.711	63.952	.000	.451
	Greenhouse-Geisser	494.132	1.418	348.472	63.952	.000	.451
Error(Target emotion group)	Sphericity Assumed	59.009	9	6.557	2.546	.008	.089
	Greenhouse-Geisser	59.009	4.254	13.871	2.546	.040	.089
Stimulus type	Sphericity Assumed	602.675	234	2.576			
	Greenhouse-Geisser	602.675	110.604	5.449			
Stimulus type* language	Sphericity Assumed	171.088	2	85.544	58.708	.000	.429
	Greenhouse-Geisser	171.088	1.269	134.771	58.708	.000	.429
Error(Stimulus type)	Sphericity Assumed	12.064	6	2.011	1.380	.226	.050
	Greenhouse-Geisser	12.064	3.808	3.168	1.380	.248	.050
Target emotion group* Stimulus type	Sphericity Assumed	227.308	156	1.457			
	Greenhouse-Geisser	227.308	99.019	2.296			
Target emotion group* Stimulus type* Language	Sphericity Assumed	111.977	6	18.663	44.358	.000	.363
	Greenhouse-Geisser	111.977	3.397	32.959	44.358	.000	.363
Error(Target emotion group* Stimulus type)	Sphericity Assumed	21.291	18	1.183	2.811	.000	.098
	Greenhouse-Geisser	21.291	10.192	2.089	2.811	.002	.098
Target emotion group * Language	Sphericity Assumed	196.905	468	.421			
	Greenhouse-Geisser	196.905	265.000	.743			

Tests of Between-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	13.479	1	13.479	8.795	.004	.101
Language	7.931	3	2.644	1.725	.169	.062
Error	119.545	78	1.533			

3 Appendix 3

Results of the simple contrasts ANOVA: ratings of the modal voice (neutral baseline) compared to the ratings of all other voice quality stimuli in the cross-language study (Chapter 4).

Apologetic-Indignant

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
stimulus	Sphericity Assumed	1273.450	14	90.961	99.569	.000	.561
	Greenhouse-Geisser	1273.450	2.968	429.027	99.569	.000	.561
stimulus * language	Sphericity Assumed	129.946	42	3.094	3.387	.000	.115
	Greenhouse-Geisser	129.946	8.905	14.593	3.387	.001	.115
Error(stimulus)	Sphericity Assumed	997.595	1092	.914			
	Greenhouse-Geisser	997.595	231.522	4.309			

Tests of Within-Subjects Contrasts

Source	stimulus	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
stimulus	modal + f0 fear vs. modal	144.912	1	144.912	64.984	.000	.454
	whispery vs. modal	222.935	1	222.935	217.393	.000	.736
	whispery + f0 fear vs. modal	352.750	1	352.750	166.467	.000	.681
	modal + f0 sadness vs. modal	96.910	1	96.910	97.732	.000	.556
	breathy vs. modal	135.069	1	135.069	167.536	.000	.682
	breathy + f0 sadness vs. modal	243.940	1	243.940	194.847	.000	.714
	modal + f0 boredom vs. modal	.965	1	.965	5.578	.021	.067
	lax-creaky vs. modal	149.936	1	149.936	74.814	.000	.490
	lax-creaky + f0 boredom vs. modal	253.080	1	253.080	181.058	.000	.699
	modal + f0 joy vs. modal	37.583	1	37.583	26.946	.000	.257
	tense vs. modal	127.257	1	127.257	214.499	.000	.733
	tense + f0 joy vs. modal	.859	1	.859	.472	.494	.006
	modal + f0 indignation vs. modal	40.801	1	40.801	17.721	.000	.185
	tense + f0 indignation vs. modal	30.715	1	30.715	11.102	.001	.125
stimulus * language	modal + f0 fear vs. modal	7.393	3	2.464	1.105	.352	.041
	whispery vs. modal	16.077	3	5.359	5.226	.002	.167
	whispery + f0 fear vs. modal	6.609	3	2.203	1.040	.380	.038
	modal + f0 sadness vs. modal	2.289	3	.763	.769	.515	.029
	breathy vs. modal	7.518	3	2.506	3.108	.031	.107
	breathy + f0 sadness vs. modal	5.077	3	1.692	1.352	.264	.049
	modal + f0 boredom vs. modal	.639	3	.213	1.232	.304	.045
	lax-creaky vs. modal	48.623	3	16.208	8.087	.000	.237
	lax-creaky + f0 boredom vs. modal	17.735	3	5.912	4.229	.008	.140
	modal + f0 joy vs. modal	9.397	3	3.132	2.246	.090	.080
	tense vs. modal	1.389	3	.463	.780	.508	.029
	tense + f0 joy vs. modal	4.006	3	1.335	.733	.535	.027
	modal + f0 indignation vs. modal	4.795	3	1.598	.694	.558	.026
	tense + f0 indignation vs. modal	4.397	3	1.466	.530	.663	.020

Error (stimulus)	modal + f0 fear vs. modal	173.937	78	2.230			
	whispery vs. modal	79.988	78	1.025			
	whispery + f0 fear vs. modal	165.285	78	2.119			
	modal + f0 sadness vs. modal	77.344	78	.992			
	breathy vs. modal	62.885	78	.806			
	breathy + f0 sadness vs. modal	97.653	78	1.252			
	modal + f0 boredom vs. modal	13.493	78	.173			
	lax-creaky vs. modal	156.322	78	2.004			
	lax-creaky + f0 boredom vs. modal	109.027	78	1.398			
	modal + f0 joy vs. modal	108.791	78	1.395			
	tense vs. modal	46.276	78	.593			
	tense + f0 joy vs. modal	142.104	78	1.822			
	modal + f0 indignation vs. modal	179.583	78	2.302			
	tense + f0 indignation vs. modal	215.799	78	2.767			

Tests of Between-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	1.526	1	1.526	18.167	.000	.189
Language	.531	3	.177	2.108	.106	.075
Error	6.550	78	.084			

Bored-Interested

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
stimulus	Sphericity Assumed	1573.959	14	112.426	133.745	.000	.632
	Greenhouse-Geisser	1573.959	3.657	430.356	133.745	.000	.632
stimulus * language	Sphericity Assumed	108.220	42	2.577	3.065	.000	.105
	Greenhouse-Geisser	108.220	10.972	9.863	3.065	.001	.105
Error(stimulus)	Sphericity Assumed	917.934	1092	.841			
	Greenhouse-Geisser	917.934	285.272	3.218			

Tests of Within-Subjects Contrasts

Source	stimulus	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
stimulus	modal + f0 fear vs. modal	6.915	1	6.915	2.797	.098	.035
	whispery vs. modal	129.418	1	129.418	81.347	.000	.511
	whispery + f0 fear vs. modal	32.165	1	32.165	9.724	.003	.111
	modal + f0 sadness vs. modal	16.241	1	16.241	11.803	.001	.131
	breathy vs. modal	97.611	1	97.611	80.442	.000	.508
	breathy + f0 sadness vs. modal	88.679	1	88.679	43.399	.000	.357
	modal + f0 boredom vs. modal	.002	1	.002	.005	.943	.000
	lax-creaky vs. modal	302.911	1	302.911	145.251	.000	.651
	lax-creaky + f0 boredom vs. modal	292.148	1	292.148	172.048	.000	.688
	modal + f0 joy vs. modal	11.448	1	11.448	6.159	.015	.073
	tense vs. modal	86.220	1	86.220	96.313	.000	.553
	tense + f0 joy vs. modal	97.026	1	97.026	73.322	.000	.485
	modal + f0 indignation vs. modal	223.789	1	223.789	128.518	.000	.622
	tense + f0 indignation vs. modal	216.307	1	216.307	126.217	.000	.618
stimulus * language	modal + f0 fear vs. modal	31.998	3	10.666	4.315	.007	.142
	whispery vs. modal	13.382	3	4.461	2.804	.045	.097
	whispery + f0 fear vs. modal	51.830	3	17.277	5.223	.002	.167
	modal + f0 sadness vs. modal	14.645	3	4.882	3.548	.018	.120
	breathy vs. modal	3.308	3	1.103	.909	.441	.034
	breathy + f0 sadness vs. modal	17.638	3	5.879	2.877	.041	.100
	modal + f0 boredom vs. modal	2.425	3	.808	2.658	.054	.093
	lax-creaky vs. modal	3.777	3	1.259	.604	.615	.023
	lax-creaky + f0 boredom vs. modal	8.524	3	2.841	1.673	.180	.060
	modal + f0 joy vs. modal	13.383	3	4.461	2.400	.074	.085
	tense vs. modal	4.163	3	1.388	1.550	.208	.056
	tense + f0 joy vs. modal	12.890	3	4.297	3.247	.026	.111
	modal + f0 indignation vs. modal	20.637	3	6.879	3.950	.011	.132
	tense + f0 indignation vs. modal	28.475	3	9.492	5.538	.002	.176

Error (stimulus)	modal + f0 fear vs. modal	192.809	78	2.472			
	whispery vs. modal	124.093	78	1.591			
	whispery + f0 fear vs. modal	258.006	78	3.308			
	modal + f0 sadness vs. modal	107.334	78	1.376			
	breathy vs. modal	94.647	78	1.213			
	breathy + f0 sadness vs. modal	159.379	78	2.043			
	modal + f0 boredom vs. modal	23.724	78	.304			
	lax-creaky vs. modal	162.664	78	2.085			
	lax-creaky + f0 boredom vs. modal	132.449	78	1.698			
	modal + f0 joy vs. modal	144.969	78	1.859			
	tense vs. modal	69.826	78	.895			
	tense + f0 joy vs. modal	103.217	78	1.323			
	modal + f0 indignation vs. modal	135.822	78	1.741			
	tense + f0 indignation vs. modal	133.675	78	1.714			

Tests of Between-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	.346	1	.346	2.315	.132	.029
Language	.437	3	.146	.973	.410	.036
Error	11.669	78	.150			

Intimate-Formal

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
stimulus	Sphericity Assumed	529.754	14	37.840	32.892	.000	.297
	Greenhouse-Geisser	529.754	2.915	181.728	32.892	.000	.297
stimulus * language	Sphericity Assumed	820.336	42	19.532	16.978	.000	.395
	Greenhouse-Geisser	820.336	8.745	93.803	16.978	.000	.395
Error(stimulus)	Sphericity Assumed	1256.267	1092	1.150			
	Greenhouse-Geisser	1256.267	227.377	5.525			

Tests of Within-Subjects Contrasts

Source	stimulus	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
stimulus	modal + f0 fear vs. modal	202.891	1	202.891	110.273	.000	.586
	whispery vs. modal	161.979	1	161.979	106.242	.000	.577
	whispery + f0 fear vs. modal	297.384	1	297.384	144.116	.000	.649
	modal + f0 sadness vs. modal	70.445	1	70.445	67.290	.000	.463
	breathy vs. modal	109.336	1	109.336	65.684	.000	.457
	breathy + f0 sadness vs. modal	192.213	1	192.213	127.280	.000	.620
	modal + f0 boredom vs. modal	5.030	1	5.030	24.015	.000	.235
	lax-creaky vs. modal	200.641	1	200.641	70.616	.000	.475
	lax-creaky + f0 boredom vs. modal	206.561	1	206.561	83.339	.000	.517
	modal + f0 joy vs. modal	82.410	1	82.410	64.655	.000	.453
	tense vs. modal	5.230	1	5.230	4.332	.041	.053
	tense + f0 joy vs. modal	73.157	1	73.157	42.181	.000	.351
	modal + f0 indignation vs. modal	90.044	1	90.044	33.488	.000	.300
	tense + f0 indignation vs. modal	124.452	1	124.452	45.411	.000	.368
stimulus * language	modal + f0 fear vs. modal	29.733	3	9.911	5.387	.002	.172
	whispery vs. modal	131.497	3	43.832	28.750	.000	.525
	whispery + f0 fear vs. modal	98.653	3	32.884	15.936	.000	.380
	modal + f0 sadness vs. modal	18.891	3	6.297	6.015	.001	.188
	breathy vs. modal	84.743	3	28.248	16.970	.000	.395
	breathy + f0 sadness vs. modal	70.258	3	23.419	15.508	.000	.374
	modal + f0 boredom vs. modal	.769	3	.256	1.224	.307	.045
	lax-creaky vs. modal	128.193	3	42.731	15.039	.000	.366
	lax-creaky + f0 boredom vs. modal	124.014	3	41.338	16.678	.000	.391
	modal + f0 joy vs. modal	12.078	3	4.026	3.159	.029	.108
	tense vs. modal	54.392	3	18.131	15.017	.000	.366
	tense + f0 joy vs. modal	37.018	3	12.339	7.115	.000	.215
	modal + f0 indignation vs. modal	93.410	3	31.137	11.580	.000	.308
	tense + f0 indignation vs. modal	76.306	3	25.435	9.281	.000	.263

Error (stimulus)	modal + f0 fear vs. modal	143.512	78	1.840			
	whispery vs. modal	118.921	78	1.525			
	whispery + f0 fear vs. modal	160.954	78	2.064			
	modal + f0 sadness vs. modal	81.657	78	1.047			
	breathy vs. modal	129.838	78	1.665			
	breathy + f0 sadness vs. modal	117.793	78	1.510			
	modal + f0 boredom vs. modal	16.336	78	.209			
	lax-creaky vs. modal	221.619	78	2.841			
	lax-creaky + f0 boredom vs. modal	193.329	78	2.479			
	modal + f0 joy vs. modal	99.420	78	1.275			
	tense vs. modal	94.173	78	1.207			
	tense + f0 joy vs. modal	135.279	78	1.734			
	modal + f0 indignation vs. modal	209.733	78	2.689			
	tense + f0 indignation vs. modal	213.764	78	2.741			

Tests of Between-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	.608	1	.608	8.434	.005	.098
Language	.208	3	.069	.964	.414	.036
Error	5.619	78	.072			

Relaxed-Stressed

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
stimulus	Sphericity Assumed	939.821	14	67.130	53.529	.000	.407
	Greenhouse-Geisser	939.821	2.673	351.563	53.529	.000	.407
stimulus * language	Sphericity Assumed	391.160	42	9.313	7.426	.000	.222
	Greenhouse-Geisser	391.160	8.020	48.774	7.426	.000	.222
Error(stimulus)	Sphericity Assumed	1369.462	1092	1.254			
	Greenhouse-Geisser	1369.462	208.515	6.568			

Tests of Within-Subjects Contrasts

Source	stimulus	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
stimulus	modal + f0 fear vs. modal	17.229	1	17.229	7.227	.009	.085
	whispery vs. modal	120.126	1	120.126	57.842	.000	.426
	whispery + f0 fear vs. modal	2.296	1	2.296	.606	.439	.008
	modal + f0 sadness vs. modal	3.319	1	3.319	2.321	.132	.029
	breathy vs. modal	128.232	1	128.232	89.088	.000	.533
	breathy + f0 sadness vs. modal	44.949	1	44.949	21.884	.000	.219
	modal + f0 boredom vs. modal	.226	1	.226	.976	.326	.012
	lax-creaky vs. modal	186.533	1	186.533	64.297	.000	.452
	lax-creaky + f0 boredom vs. modal	179.434	1	179.434	72.171	.000	.481
	modal + f0 joy vs. modal	.041	1	.041	.028	.867	.000
	tense vs. modal	86.721	1	86.721	80.363	.000	.507
	tense + f0 joy vs. modal	31.158	1	31.158	15.659	.000	.167
	modal + f0 indignation vs. modal	87.475	1	87.475	33.628	.000	.301
	tense + f0 indignation vs. modal	85.561	1	85.561	31.357	.000	.287
stimulus * language	modal + f0 fear vs. modal	47.375	3	15.792	6.624	.000	.203
	whispery vs. modal	50.204	3	16.735	8.058	.000	.237
	whispery + f0 fear vs. modal	66.708	3	22.236	5.870	.001	.184
	modal + f0 sadness vs. modal	24.077	3	8.026	5.612	.002	.178
	breathy vs. modal	33.732	3	11.244	7.812	.000	.231
	breathy + f0 sadness vs. modal	38.372	3	12.791	6.227	.001	.193
	modal + f0 boredom vs. modal	.409	3	.136	.588	.624	.022
	lax-creaky vs. modal	82.444	3	27.481	9.473	.000	.267
	lax-creaky + f0 boredom vs. modal	67.645	3	22.548	9.069	.000	.259
	modal + f0 joy vs. modal	21.481	3	7.160	4.893	.004	.158
	tense vs. modal	14.204	3	4.735	4.387	.007	.144
	tense + f0 joy vs. modal	32.605	3	10.868	5.462	.002	.174
	modal + f0 indignation vs. modal	55.776	3	18.592	7.147	.000	.216
	tense + f0 indignation vs. modal	48.147	3	16.049	5.882	.001	.184

Error (stimulus)	modal + f0 fear vs. modal	185.948	78	2.384			
	whispery vs. modal	161.991	78	2.077			
	whispery + f0 fear vs. modal	295.447	78	3.788			
	modal + f0 sadness vs. modal	111.541	78	1.430			
	breathy vs. modal	112.272	78	1.439			
	breathy + f0 sadness vs. modal	160.208	78	2.054			
	modal + f0 boredom vs. modal	18.095	78	.232			
	lax-creaky vs. modal	226.285	78	2.901			
	lax-creaky + f0 boredom vs. modal	193.925	78	2.486			
	modal + f0 joy vs. modal	114.136	78	1.463			
	tense vs. modal	84.171	78	1.079			
	tense + f0 joy vs. modal	155.204	78	1.990			
	modal + f0 indignation vs. modal	202.900	78	2.601			
	tense + f0 indignation vs. modal	212.828	78	2.729			

Tests of Between-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	.080	1	.080	.900	.346	.011
Language	.541	3	.180	2.032	.116	.072
Error	6.925	78	.089			

Sad-Happy

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
stimulus	Sphericity Assumed	832.599	14	59.471	76.949	.000	.497
	Greenhouse-Geisser	832.599	2.441	341.064	76.949	.000	.497
stimulus * language	Sphericity Assumed	47.501	42	1.131	1.463	.030	.053
	Greenhouse-Geisser	47.501	7.324	6.486	1.463	.179	.053
Error(stimulus)	Sphericity Assumed	843.971	1092	.773			
	Greenhouse-Geisser	843.971	190.412	4.432			

Tests of Within-Subjects Contrasts

Source	stimulus	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
stimulus	modal + f0 fear vs. modal	22.974	1	22.974	15.389	.000	.165
	whispery vs. modal	109.517	1	109.517	126.126	.000	.618
	whispery + f0 fear vs. modal	137.978	1	137.978	92.653	.000	.543
	modal + f0 sadness vs. modal	29.678	1	29.678	51.113	.000	.396
	breathy vs. modal	61.242	1	61.242	94.333	.000	.547
	breathy + f0 sadness vs. modal	116.667	1	116.667	139.010	.000	.641
	modal + f0 boredom vs. modal	.292	1	.292	2.587	.112	.032
	lax-creaky vs. modal	195.484	1	195.484	126.624	.000	.619
	lax-creaky + f0 boredom vs. modal	226.441	1	226.441	188.909	.000	.708
	modal + f0 joy vs. modal	4.425	1	4.425	5.173	.026	.062
	tense vs. modal	35.345	1	35.345	79.193	.000	.504
	tense + f0 joy vs. modal	14.654	1	14.654	11.578	.001	.129
	modal + f0 indignation vs. modal	57.440	1	57.440	29.494	.000	.274
	tense + f0 indignation vs. modal	42.934	1	42.934	19.984	.000	.204
stimulus * language	modal + f0 fear vs. modal	8.967	3	2.989	2.002	.121	.072
	whispery vs. modal	12.527	3	4.176	4.809	.004	.156
	whispery + f0 fear vs. modal	18.742	3	6.247	4.195	.008	.139
	modal + f0 sadness vs. modal	6.368	3	2.123	3.655	.016	.123
	breathy vs. modal	7.259	3	2.420	3.727	.015	.125
	breathy + f0 sadness vs. modal	8.496	3	2.832	3.375	.023	.115
	modal + f0 boredom vs. modal	.427	3	.142	1.261	.294	.046
	lax-creaky vs. modal	8.047	3	2.682	1.738	.166	.063
	lax-creaky + f0 boredom vs. modal	9.728	3	3.243	2.705	.051	.094
	modal + f0 joy vs. modal	1.554	3	.518	.605	.613	.023
	tense vs. modal	2.027	3	.676	1.514	.218	.055
	tense + f0 joy vs. modal	3.017	3	1.006	.795	.501	.030
	modal + f0 indignation vs. modal	.367	3	.122	.063	.979	.002
	tense + f0 indignation vs. modal	2.510	3	.837	.389	.761	.015

Error (stimulus)	modal + f0 fear vs. modal	116.441	78	1.493			
	whispery vs. modal	67.728	78	.868			
	whispery + f0 fear vs. modal	116.157	78	1.489			
	modal + f0 sadness vs. modal	45.290	78	.581			
	breathy vs. modal	50.639	78	.649			
	breathy + f0 sadness vs. modal	65.463	78	.839			
	modal + f0 boredom vs. modal	8.808	78	.113			
	lax-creaky vs. modal	120.417	78	1.544			
	lax-creaky + f0 boredom vs. modal	93.497	78	1.199			
	modal + f0 joy vs. modal	66.730	78	.856			
	tense vs. modal	34.812	78	.446			
	tense + f0 joy vs. modal	98.723	78	1.266			
	modal + f0 indignation vs. modal	151.906	78	1.948			
	tense + f0 indignation vs. modal	167.581	78	2.148			

Tests of Between-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	13.062	1	13.062	59.678	.000	.433
Language	.386	3	.129	.588	.624	.022
Error	17.072	78	.219			

Scared-Fearless

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
stimulus	Sphericity Assumed	870.926	14	62.209	56.945	.000	.422
	Greenhouse-Geisser	870.926	2.287	380.787	56.945	.000	.422
stimulus * language	Sphericity Assumed	109.284	42	2.602	2.382	.000	.084
	Greenhouse-Geisser	109.284	6.862	15.927	2.382	.025	.084
Error(stimulus)	Sphericity Assumed	1192.953	1092	1.092			
	Greenhouse-Geisser	1192.953	178.399	6.687			

Tests of Within-Subjects Contrasts

Source	stimulus	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
stimulus	modal + f0 fear vs. modal	368.590	1	368.590	271.848	.000	.777
	whispery vs. modal	167.610	1	167.610	171.747	.000	.688
	whispery + f0 fear vs. modal	565.105	1	565.105	557.737	.000	.877
	modal + f0 sadness vs. modal	121.964	1	121.964	158.637	.000	.670
	breathy vs. modal	68.326	1	68.326	76.850	.000	.496
	breathy + f0 sadness vs. modal	263.244	1	263.244	357.953	.000	.821
	modal + f0 boredom vs. modal	.991	1	.991	4.960	.029	.060
	lax-creaky vs. modal	43.850	1	43.850	20.882	.000	.211
	lax-creaky + f0 boredom vs. modal	114.604	1	114.604	74.441	.000	.488
	modal + f0 joy vs. modal	122.620	1	122.620	106.077	.000	.576
	tense vs. modal	51.277	1	51.277	56.120	.000	.418
	tense + f0 joy vs. modal	39.152	1	39.152	21.242	.000	.214
	modal + f0 indignation vs. modal	34.529	1	34.529	10.840	.001	.122
	tense + f0 indignation vs. modal	44.768	1	44.768	12.616	.001	.139
stimulus * language	modal + f0 fear vs. modal	6.768	3	2.256	1.664	.182	.060
	whispery vs. modal	29.266	3	9.755	9.996	.000	.278
	whispery + f0 fear vs. modal	15.135	3	5.045	4.979	.003	.161
	modal + f0 sadness vs. modal	1.852	3	.617	.803	.496	.030
	breathy vs. modal	20.254	3	6.751	7.594	.000	.226
	breathy + f0 sadness vs. modal	8.667	3	2.889	3.928	.011	.131
	modal + f0 boredom vs. modal	.255	3	.085	.426	.735	.016
	lax-creaky vs. modal	7.263	3	2.421	1.153	.333	.042
	lax-creaky + f0 boredom vs. modal	7.459	3	2.486	1.615	.193	.058
	modal + f0 joy vs. modal	.236	3	.079	.068	.977	.003
	tense vs. modal	.376	3	.125	.137	.938	.005
	tense + f0 joy vs. modal	4.631	3	1.544	.838	.477	.031
	modal + f0 indignation vs. modal	21.328	3	7.109	2.232	.091	.079
	tense + f0 indignation vs. modal	12.071	3	4.024	1.134	.341	.042

Error (stimulus)	modal + f0 fear vs. modal	105.758	78	1.356			
	whispery vs. modal	76.121	78	.976			
	whispery + f0 fear vs. modal	79.030	78	1.013			
	modal + f0 sadness vs. modal	59.968	78	.769			
	breathy vs. modal	69.348	78	.889			
	breathy + f0 sadness vs. modal	57.362	78	.735			
	modal + f0 boredom vs. modal	15.577	78	.200			
	lax-creaky vs. modal	163.795	78	2.100			
	lax-creaky + f0 boredom vs. modal	120.084	78	1.540			
	modal + f0 joy vs. modal	90.164	78	1.156			
	tense vs. modal	71.268	78	.914			
	tense + f0 joy vs. modal	143.764	78	1.843			
	modal + f0 indignation vs. modal	248.451	78	3.185			
	tense + f0 indignation vs. modal	276.787	78	3.549			

Tests of Between-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	.108	1	.108	.887	.349	.011
Language	1.593	3	.531	4.362	.007	.144
Error	9.496	78	.122			

4 Appendix 4

Shown is the raters' percent agreement in each of the subtests in the cross-language study (Chapter 4).

The data shows percentage of the raters that assigned ratings above +/- 1 to a particular voice quality stimulus. Ratings within the range of -1 to +1 are interpreted here as suggestive of 'no affect' signalled. Cases in which 75% of the raters or more agreed in voice to affect association are shown in bold type.

Table 4.1 *Raters' percent agreement in the cross-language study. Cases of 75% agreement and above in voice to affect association are shown in bold type.*

apologetic-indignant	apologetic				indignant				no affect			
	HE	R	S	J	HE	R	S	J	HE	R	S	J
modal + f_0 'fear'	55	71	65	**90**	20	10	35	10	25	19	0	0
whispery	**80**	**90**	**90**	52	0	5	0	24	20	5	10	24
whispery + f_0 'fear'	**90**	**100**	**85**	**95**	5	0	15	5	5	0	0	0
modal + f_0 'sadness'	50	52	70	**81**	10	19	15	10	40	29	15	10
breathy	55	**81**	70	67	15	5	15	29	30	14	15	5
breathy + f_0 'sadness'	**80**	**90**	65	67	0	0	5	24	20	10	30	10
modal + f_0 'boredom'	0	10	15	10	50	67	60	71	50	24	25	19
lax-creaky	60	**86**	70	33	15	5	10	48	25	10	20	19
lax-creaky + f_0 'boredom'	**75**	**100**	**90**	43	5	0	5	33	20	0	5	24
modal + f_0 'joy'	30	10	45	**86**	30	43	40	5	40	43	15	10
tense	0	5	10	5	**90**	**95**	**90**	**81**	10	0	0	14
tense + f_0 'joy'	10	0	25	38	60	**76**	70	52	30	24	5	10
modal + f_0 'indignation'	10	0	20	29	**90**	**95**	**80**	71	0	5	0	0
tense + f_0 'indignation'	10	0	25	29	**85**	**100**	**75**	67	5	0	0	5
modal	0	19	25	24	40	62	50	48	60	19	25	29
bored-interested	**bored**				**interested**				**no affect**			
	HE	R	S	J	HE	R	S	J	HE	R	S	J
modal + f_0 'fear'	25	24	50	14	50	48	45	67	25	29	5	19
whispery	65	**90**	**90**	**81**	10	5	0	5	25	5	10	14
whispery + f_0 'fear'	55	**86**	**75**	43	15	10	25	29	30	5	0	29
modal + f_0 'sadness'	60	38	60	67	10	24	30	14	30	38	10	19
breathy	**75**	**86**	**90**	**81**	10	5	5	10	15	10	5	10
breathy + f_0 'sadness'	**90**	**90**	**80**	**76**	5	5	15	5	5	5	5	19
modal + f_0 'boredom'	25	10	50	71	30	67	35	24	40	24	15	5
lax-creaky	**80**	**90**	**95**	**95**	5	10	0	5	15	0	5	0
lax-creaky + f_0 'boredom'	**85**	**90**	**90**	**90**	10	5	0	5	5	5	10	5
modal + f_0 'joy'	20	24	30	14	45	57	45	67	35	19	25	19
tense	15	5	25	29	65	**90**	70	57	20	5	5	14
tense + f_0 'joy'	10	0	10	5	**85**	**100**	**90**	**90**	5	0	0	5
modal + f_0 'indignation'	5	5	15	0	**95**	**95**	**85**	**100**	0	0	0	0
tense + f_0 'indignation'	5	5	10	0	**95**	**90**	**90**	**100**	0	5	0	0
modal	20	19	45	57	40	71	40	24	40	10	15	19

Table 4.2. *Raters' percent agreement in the cross-language study. Cases of 75% agreement and above in voice to affect association are shown in bold type.*

intimate-formal	intimate				formal				no affect			
	HE	R	S	J	HE	R	S	J	HE	R	S	J
modal + f_0 'fear'	45	67	**90**	57	30	14	5	38	25	19	5	5
whispery	**80**	**90**	25	14	10	0	70	57	10	10	5	29
whispery + f_0 'fear'	**90**	**86**	**75**	43	0	0	15	48	10	14	10	10
modal + f_0 'sadness'	50	33	45	29	20	52	25	48	30	14	30	24
breathy	**75**	**81**	25	33	20	10	65	62	5	10	10	5
breathy + f_0 'sadness'	**80**	**81**	50	38	5	0	15	48	15	19	35	14
modal + f_0 'boredom'	0	0	5	14	**75**	**95**	**90**	57	25	5	5	29
lax-creaky	**80**	**81**	45	24	20	5	50	57	0	14	5	19
lax-creaky + f_0 'boredom'	**90**	**86**	35	24	10	0	50	52	0	14	15	24
modal + f_0 'joy'	45	38	55	38	15	52	25	33	40	10	20	29
tense	0	0	20	52	**90**	**100**	60	43	10	0	20	5
tense + f_0 'joy'	25	5	65	**76**	65	**95**	25	19	10	0	10	5
modal + f_0 'indignation'	20	5	**75**	**86**	**75**	**90**	25	10	5	5	0	5
tense + f_0 'indignation'	25	10	**75**	**90**	60	**86**	15	5	15	5	10	5
modal	0	0	10	33	65	**90**	85	52	35	10	5	14
relaxed-stressed	**relaxed**				**stressed**				**no affect**			
	HE	R	S	J	HE	R	S	J	HE	R	S	J
modal + f_0 'fear'	15	38	10	38	65	33	**80**	48	20	29	10	14
whispery	**80**	**95**	**100**	52	5	0	0	38	10	0	0	10
whispery + f_0 'fear'	35	**81**	35	24	55	14	60	52	10	5	5	24
modal + f_0 'sadness'	30	57	40	33	15	19	10	48	55	24	50	19
breathy	**85**	**76**	**95**	52	0	10	0	43	15	14	5	5
breathy + f_0 'sadness'	65	**81**	**80**	38	10	5	0	48	25	14	20	14
modal + f_0 'boredom'	10	19	55	38	35	62	35	43	55	19	10	19
lax-creaky	**85**	**86**	**95**	38	5	14	5	57	10	0	0	5
lax-creaky + f_0 'boredom'	**95**	**81**	**95**	43	5	10	5	57	0	10	0	0
modal + f_0 'joy'	25	38	15	48	40	29	45	29	35	33	40	24
tense	0	10	15	33	**85**	**90**	70	62	15	0	15	5
tense + f_0 'joy'	10	10	5	48	**80**	**81**	**85**	48	10	10	10	5
modal + f_0 'indignation'	5	5	5	52	**95**	**95**	**95**	43	0	0	0	5
tense + f_0 'indignation'	5	5	5	52	**90**	**95**	**95**	48	5	0	0	0
modal	20	14	50	48	50	62	40	33	30	24	10	19

Table 4.3. *Raters' percent agreement in the cross-language study. Cases of 75% agreement and above in voice to affect association are shown in bold type.*

sad-happy	sad				happy				no affect			
	HE	R	S	J	HE	R	S	J	HE	R	S	J
modal + f_0 'fear'	50	71	60	62	40	24	20	24	10	5	20	14
whispery	**90**	**95**	**80**	67	0	0	0	5	10	5	20	29
whispery + f_0 'fear'	**80**	**95**	**90**	**76**	10	5	5	10	10	0	5	14
modal + f_0 'sadness'	65	71	55	57	0	10	20	14	35	19	25	29
breathy	**75**	**76**	**75**	67	5	0	5	19	20	24	20	14
breathy + f_0 'sadness'	**85**	**90**	**85**	**86**	5	0	10	10	10	10	5	5
modal + f_0 'boredom'	15	14	30	29	5	67	30	24	80	19	40	48
lax-creaky	**80**	**81**	**85**	62	5	14	5	5	15	5	10	33
lax-creaky + f_0 'boredom'	**90**	**86**	**100**	**81**	5	0	0	5	5	14	0	14
modal + f_0 'joy'	45	24	60	33	20	48	40	24	35	29	0	43
tense	5	19	25	14	50	**81**	50	52	45	0	25	33
tense + f_0 'joy'	25	14	30	43	60	**76**	60	57	15	10	10	0
modal + f_0 'indignation'	15	14	25	19	**75**	**86**	70	**76**	5	0	5	5
tense + f_0 'indignation'	30	29	30	24	65	71	70	76	5	0	0	0
modal	15	14	35	33	15	52	30	10	70	33	35	57
scared-fearless	**scared**				**fearless**				**no affect**			
	HE	R	S	J	HE	R	S	J	HE	R	S	J
modal + f_0 'fear'	70	**86**	**80**	**86**	10	5	5	0	20	10	15	14
whispery	60	71	40	43	10	14	40	24	30	14	20	33
whispery + f_0 'fear'	**95**	**95**	**90**	**90**	0	0	5	0	5	5	5	10
modal + f_0 'sadness'	30	43	30	67	30	29	40	5	40	29	30	29
breathy	25	48	35	38	35	24	55	43	40	29	10	19
breathy + f_0 'sadness'	**80**	**76**	45	71	0	5	15	5	20	19	40	24
modal + f_0 'boredom'	0	0	5	5	65	**81**	**80**	62	35	19	15	33
lax-creaky	15	43	40	33	40	38	45	48	45	19	15	19
lax-creaky + f_0 'boredom'	50	57	45	43	25	29	40	29	25	14	15	29
modal + f_0 'joy'	45	19	45	62	20	57	15	14	35	24	40	24
tense	0	10	15	10	**90**	**90**	**85**	**86**	10	0	0	5
tense + f_0 'joy'	30	5	45	48	60	**76**	45	43	10	19	10	10
modal + f_0 'indignation'	35	24	70	43	60	**76**	25	57	5	0	5	0
tense + f_0 'indignation'	40	24	60	43	55	71	30	57	5	5	10	0
modal	0	0	5	5	60	**90**	**75**	67	40	10	20	29

5 Appendix 5

Raters' agreement in the cross-language study (Chapter 4).

For each of the language groups tested in the cross-language study, Pearson's *r* correlation coefficients were calculated showing correlation of each individual rater with the group mean for a particular subtest. This allows one to ascertain what raters within each group of listeners were found to assign ratings very different to those of a group as a whole.

These coefficients are presented in Tables 5.1 - 5.3 below. Cases of Pearson's *r* below 0.5 (suggestive of substantial difference between individual raters and a group as a whole) are shown in bold type.

Table 5.1. *Pearson's* r *showing correlation between the ratings of individual raters and the group mean.*

	Apologetic-Indignant				Bored-Interested			
	HE	**R**	**S**	**J**	**HE**	**R**	**S**	**J**
S1	0.82	0.97	0.96	0.65	0.92	0.92	0.88	0.99
S2	0.96	0.90	**0.25**	0.88	**-0.22**	0.96	0.61	0.95
S3	**0.35**	0.93	0.92	0.70	0.91	0.95	0.95	0.94
S4	0.92	0.87	0.67	**0.41**	0.92	0.64	0.85	0.96
S5	0.93	0.96	0.10	0.83	0.74	0.94	0.66	0.83
S6	0.97	0.92	**0.35**	**0.39**	0.94	0.90	0.95	0.89
S7	0.96	0.94	0.55	0.92	0.91	0.90	**0.41**	0.95
S8	0.90	0.97	0.79	0.66	0.89	0.98	0.66	0.92
S9	0.96	0.93	0.78	0.81	0.97	0.94	0.73	0.96
S10	0.97	0.81	0.97	**0.42**	0.94	0.78	0.83	0.62
S11	**0.17**	0.91	0.69	0.67	**-0.52**	0.97	0.73	0.93
S12	0.89	0.97	**-0.11**	0.65	0.97	0.97	0.55	0.91
S13	0.91	**0.49**	0.91	0.73	**0.27**	0.87	0.79	0.85
S14	**0.33**	0.95	0.97	0.86	0.65	0.60	0.92	0.97
S15	0.94	0.92	0.94	0.91	0.93	0.94	0.98	0.95
S16	0.77	0.98	0.95	0.79	0.97	0.98	0.95	0.98
S17	0.91	0.92	0.94	0.85	0.98	0.95	0.82	**-0.38**
S18	0.96	0.79	0.90	**-0.17**	0.94	0.94	0.91	0.80
S19	0.96	0.95	0.66	0.87	0.89	0.96	0.67	0.97
S20	0.97	0.97	**0.40**	0.86	0.93	**-0.52**	0.72	0.86
S21		0.91		0.92		0.96		0.93
	N=20	N=21	N=20	N=21	N=20	N=21	N=20	N=21

Table 5.2. *Pearson's* r *showing correlation between the ratings of individual raters and the group mean.*

	Intimate-Formal				Relaxed-Stressed			
	HE	**R**	**S**	**J**	**HE**	**R**	**S**	**J**
S1	0.72	0.95	0.76	0.87	0.93	0.97	0.98	0.50
S2	0.78	0.93	0.77	0.89	0.95	0.96	0.78	**-0.13**
S3	0.95	0.90	0.97	0.50	0.74	0.92	0.98	**-0.14**
S4	0.76	0.90	0.80	0.85	0.66	**0.36**	0.96	**0.27**
S5	0.89	0.90	0.97	**-0.06**	0.97	0.91	0.72	**0.25**
S6	**0.44**	0.96	**0.28**	0.75	0.96	0.85	0.92	**-0.15**
S7	0.96	0.96	**0.43**	0.75	0.84	0.90	0.57	**0.27**
S8	0.88	0.92	0.81	0.73	0.74	0.94	0.96	**-0.02**
S9	0.89	0.95	0.78	**0.00**	0.93	0.94	0.77	**0.10**
S10	0.91	**0.29**	0.58	**0.34**	0.91	0.78	0.91	**0.26**
S11	**0.13**	0.91	0.52	0.84	0.86	0.99	0.89	**0.00**
S12	0.84	0.94	0.91	0.75	0.94	0.98	0.98	**0.05**
S13	0.60	0.94	**0.41**	0.78	0.81	0.84	0.78	**0.10**
S14	**0.09**	0.88	0.57	0.73	0.75	0.51	0.84	**0.11**
S15	0.85	0.86	0.69	**0.07**	0.93	0.98	0.89	**0.20**
S16	0.74	0.95	**-0.09**	0.78	**-0.76**	0.98	0.88	**0.18**
S17	0.91	0.84	**0.00**	**-0.72**	0.62	0.96	0.92	**0.06**
S18	0.96	0.96	**-0.18**	**-0.12**	0.82	0.79	0.87	**0.44**
S19	0.98	0.98	0.61	0.74	0.99	0.96	0.89	**0.25**
S20	0.99	0.94	0.81	0.87	0.86	0.85	0.88	**0.16**
S21		0.98		0.85		**-0.94**		**0.22**
	N=20	N=21	N=20	N=21	N=20	N=21	N=20	N=21

Table 5.3. *Pearson's r showing correlation between the ratings of individual raters and the group mean.*

	S-H				F-S			
	HE	**R**	**S**	**J**	**HE**	**R**	**S**	**J**
S1	0.80	0.83	0.94	0.92	0.67	0.88	0.83	0.89
S2	**-0.59**	0.93	0.89	0.95	0.59	0.81	0.75	0.53
S3	0.57	0.93	0.91	0.62	0.90	0.90	0.67	0.57
S4	0.88	**0.42**	0.79	0.87	0.82	**0.48**	**0.24**	0.83
S5	0.89	0.96	**-0.06**	**-0.69**	0.88	0.94	0.89	0.83
S6	0.90	0.92	**-0.36**	0.74	0.52	0.91	0.91	0.61
S7	0.87	0.90	**-0.01**	**0.31**	0.95	0.56	**0.24**	0.66
S8	0.97	**0.18**	0.60	**0.35**	0.78	0.87	0.75	0.86
S9	0.98	0.95	0.65	**-0.21**	0.88	0.92	0.66	0.56
S10	0.89	0.58	0.94	**0.01**	0.94	0.58	0.93	0.79
S11	**-0.37**	**0.01**	0.78	0.94	0.53	0.92	0.51	**-0.06**
S12	0.94	**0.47**	**-0.45**	0.95	0.73	0.85	0.84	0.60
S13	**0.39**	0.63	0.96	0.99	0.68	0.66	**0.19**	0.65
S14	**0.18**	0.96	0.98	0.92	**0.34**	0.88	0.73	0.64
S15	0.96	0.96	0.95	0.65	0.84	0.84	0.70	0.78
S16	0.93	0.97	0.97	0.94	**0.48**	0.93	**0.33**	0.74
S17	0.98	0.96	0.90	**0.30**	0.67	0.94	0.84	0.46
S18	0.99	0.78	0.96	0.86	0.77	0.92	**0.22**	**0.01**
S19	**0.47**	0.98	**-0.60**	0.83	0.70	**-0.19**	0.77	**0.42**
S20	0.95	0.94	0.61	0.92	0.90	0.82	0.62	0.75
S21		0.79		0.97		0.82		0.72
	N=20	N=21	N=20	N=21	N=20	N=21	N=20	N=21

6 Appendix 6

Raters' agreement in the cross-language study (Chapter 4).

Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for each individual rater within a language group, for each stimulus, for each subtest. Each dot in the plots corresponds to the data for one rater. Group mean and standard deviation (for a particular language group) are shown as a black dot.

The data allows one to ascertain the extent of agreement/disagreement the raters demonstrated in rating a particular stimulus.

The data also allows one to infer the degree of each rater's confidence in voice to affect association. Given the range of the rating scale, standard deviation of 1 and below (across 10 randomisations) suggests relatively confident voice to affect association.

In the figures, the scatter plots are arranged so as to allow comparison across the three types of stimuli within each of the 'target emotion group' ('fear', 'sadness', 'boredom', 'joy', 'indignation'[28]) in the following sequence:

'VQ only'

'f_0 only'

'VQ + f_0'

[28] See also Table 3.2 in Chapter 3.

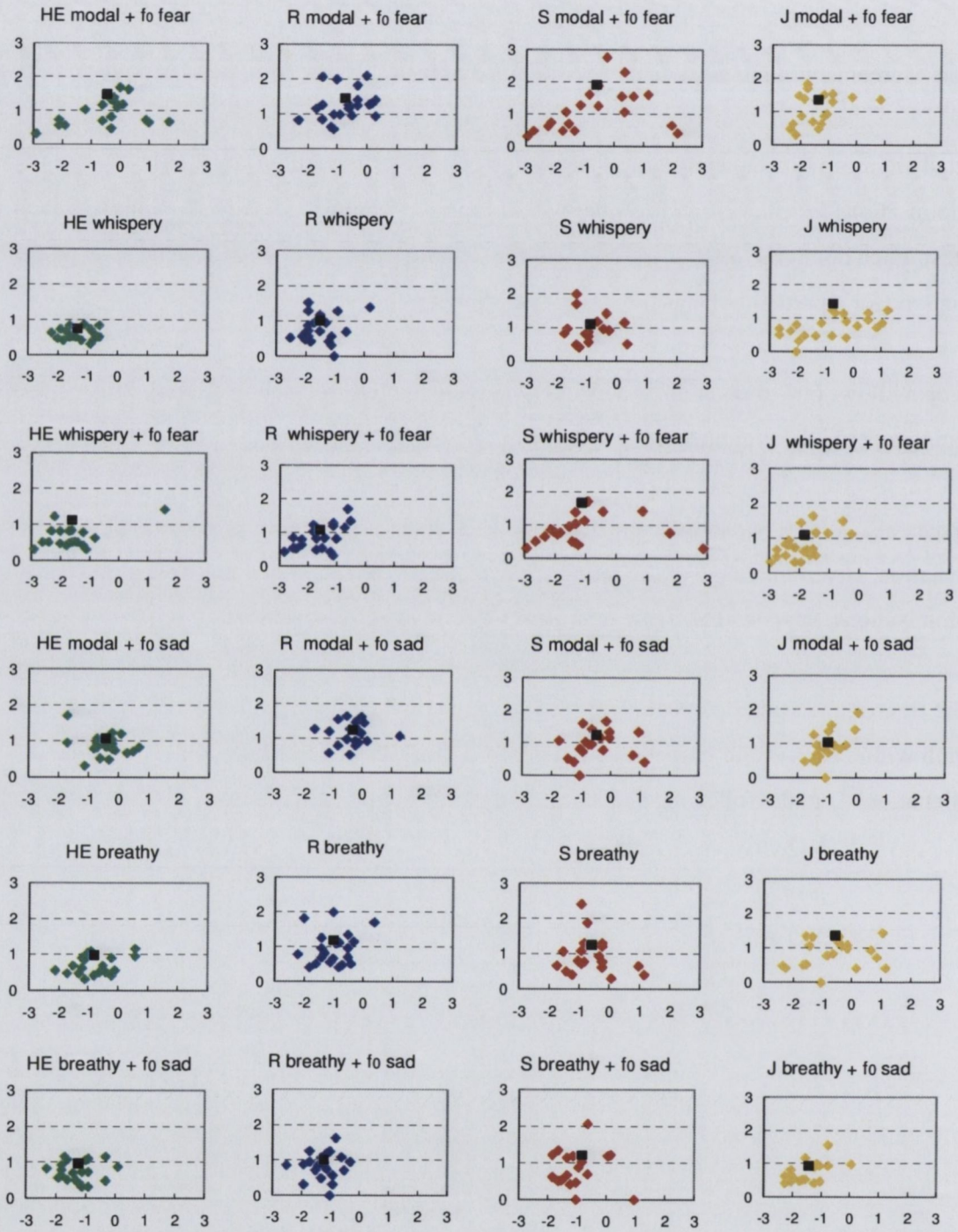


Figure 6.1. *Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for individual raters. Each dot in the plots corresponds to the data for one rater. Standard deviation of 1 and below suggests confident voice to affect association. Group mean and standard deviation are shown as a black dot.*

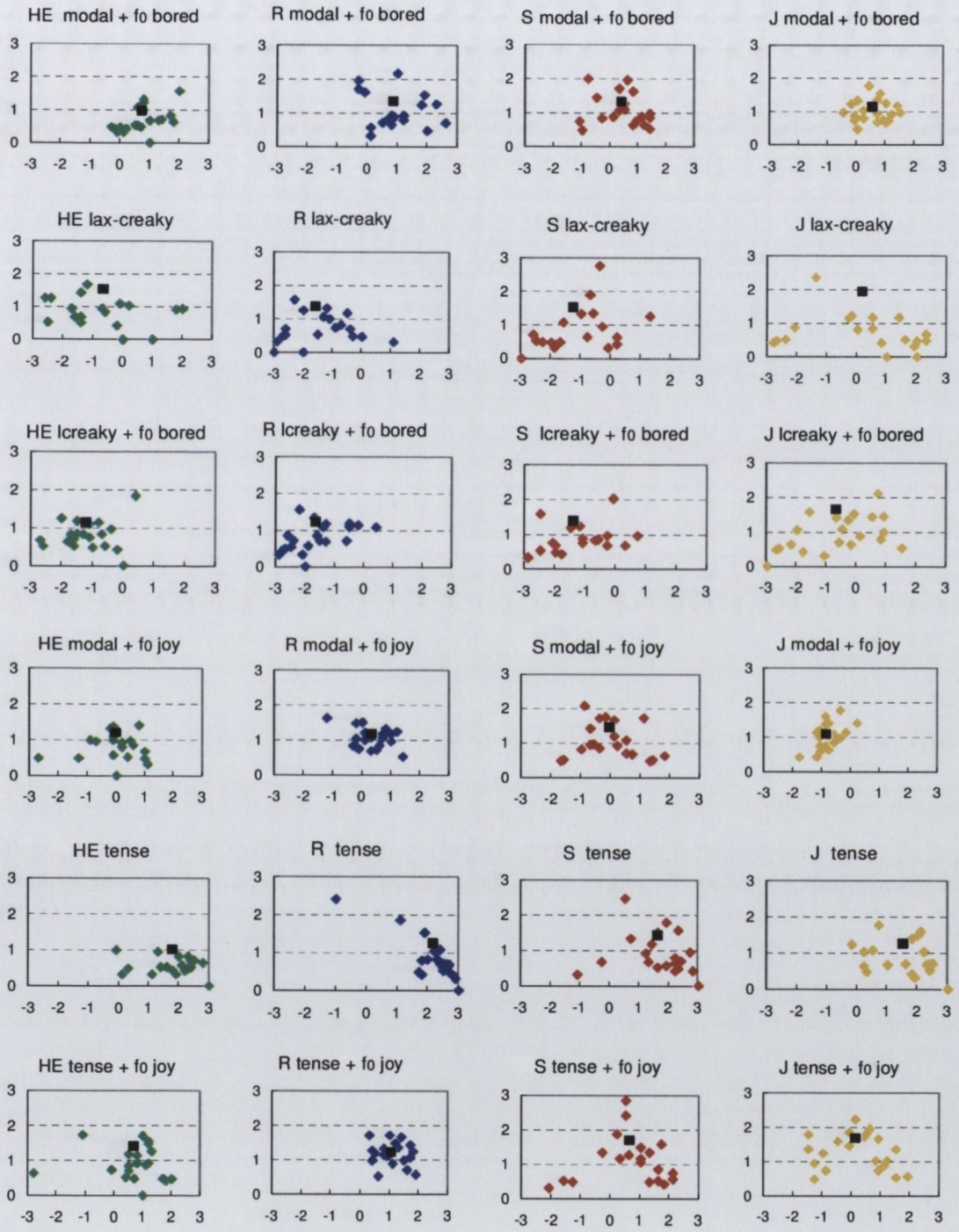


Figure 6.2. *Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for individual raters. Each dot in the plots corresponds to the data for one rater. Standard deviation of 1 and below suggests confident voice to affect association. Group mean and standard deviation are shown as a black dot.*

APLOGETIC-INDIGNANT (CONT)

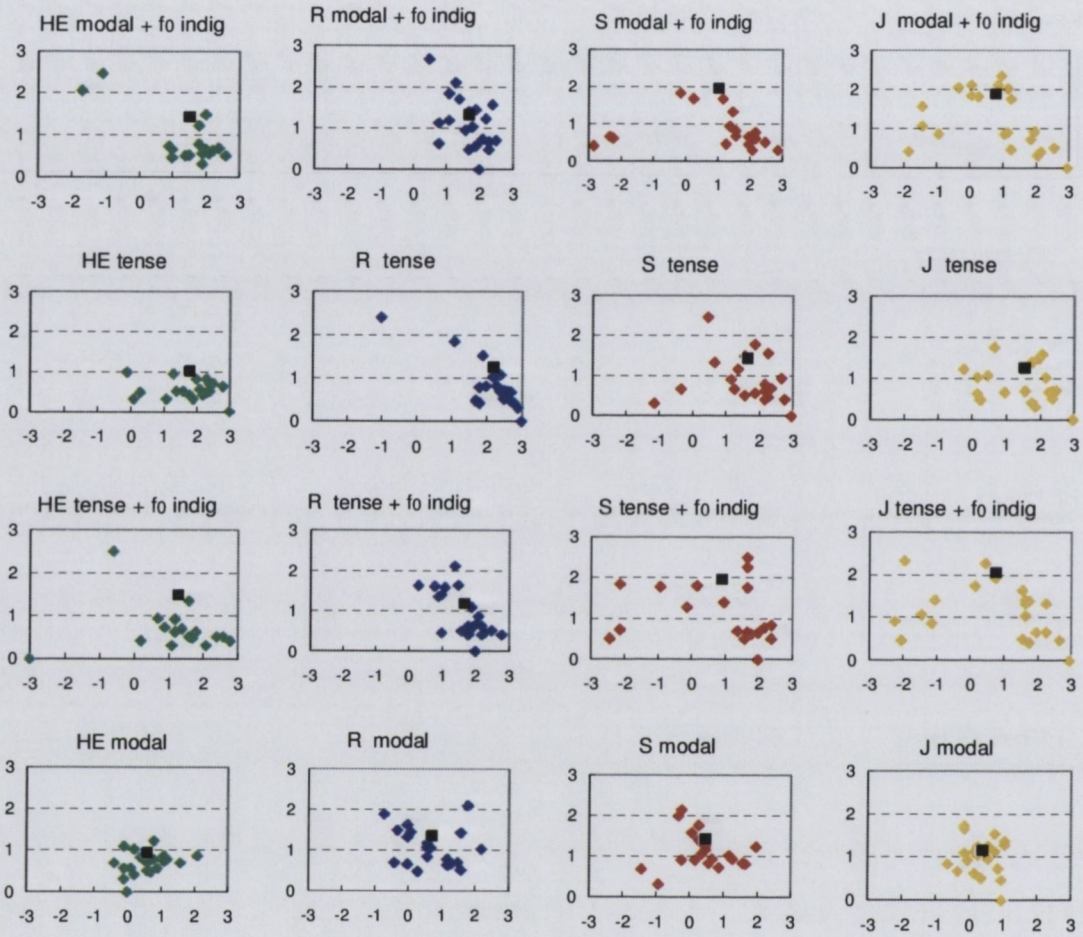


Figure 6.3. *Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for individual raters. Each dot in the plots corresponds to the data for one rater. Standard deviation of 1 and below suggests confident voice to affect association. Group mean and standard deviation are shown as a black dot.*

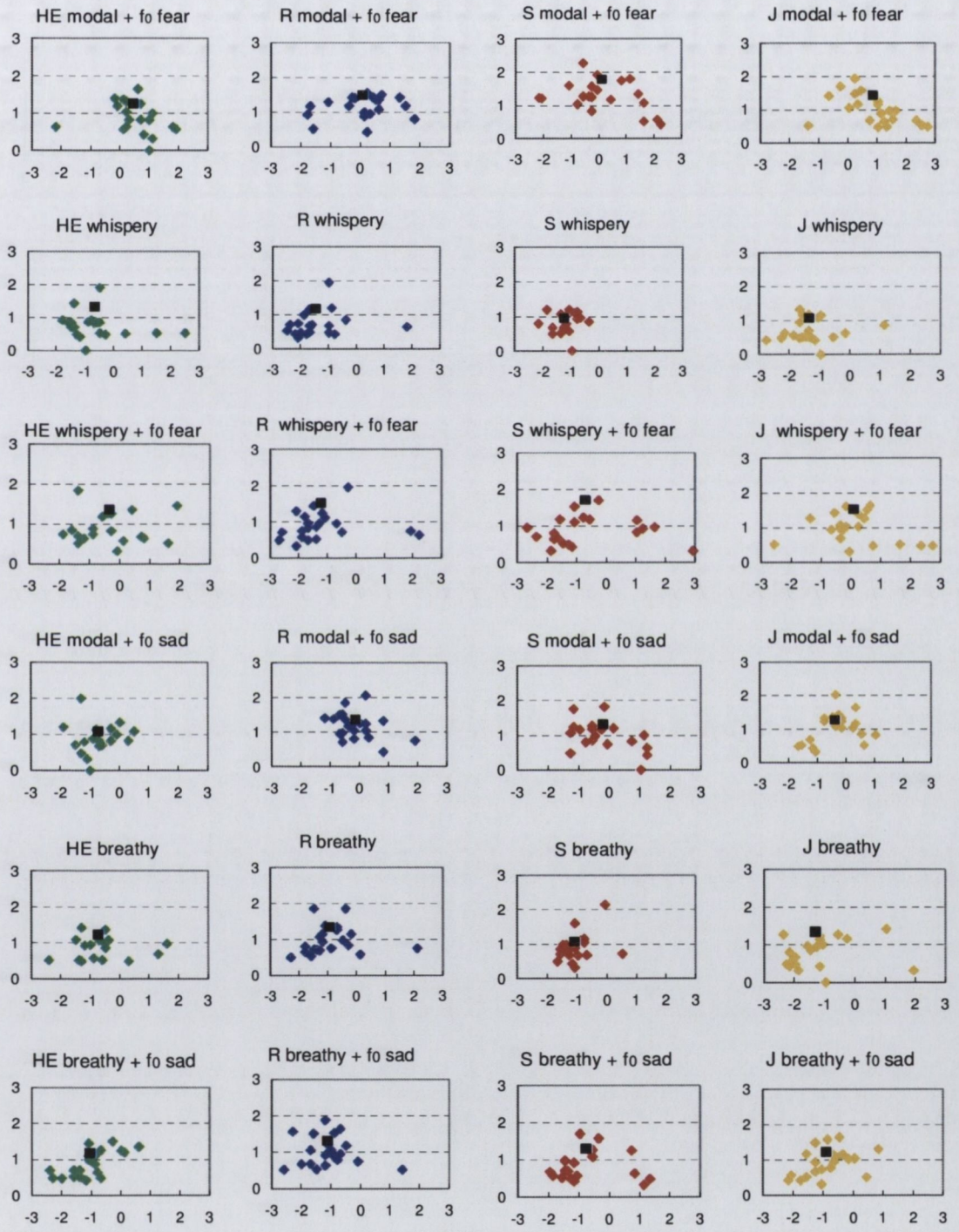


Figure 6.4. *Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for individual raters. Each dot in the plots corresponds to the data for one rater. Standard deviation of 1 and below suggests confident voice to affect association. Group mean and standard deviation are shown as a black dot.*

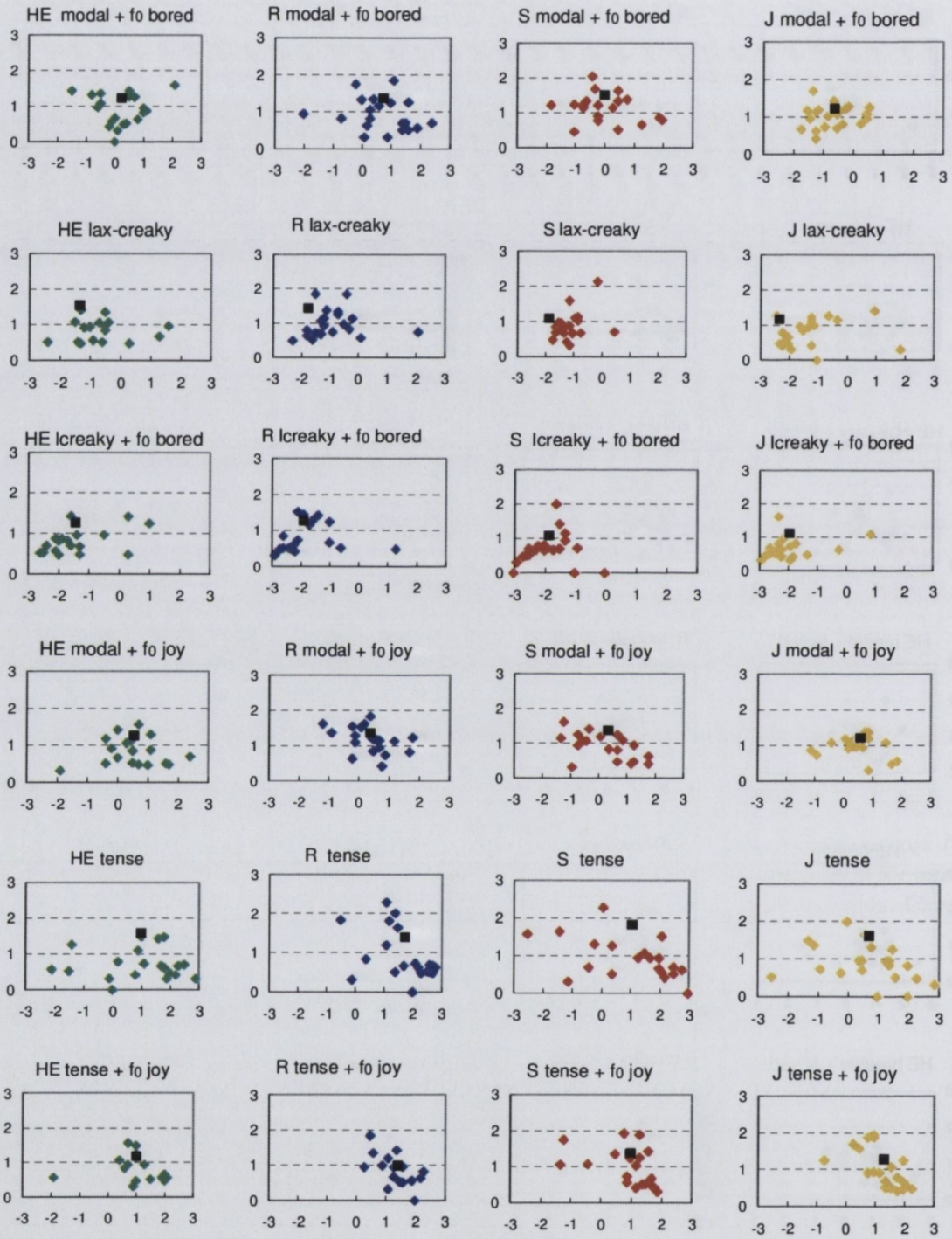


Figure 6.5. *Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for individual raters. Each dot in the plots corresponds to the data for one rater. Standard deviation of 1 and below suggests confident voice to affect association. Group mean and standard deviation are shown as a black dot.*

BORED-INTERESTED (CONT)

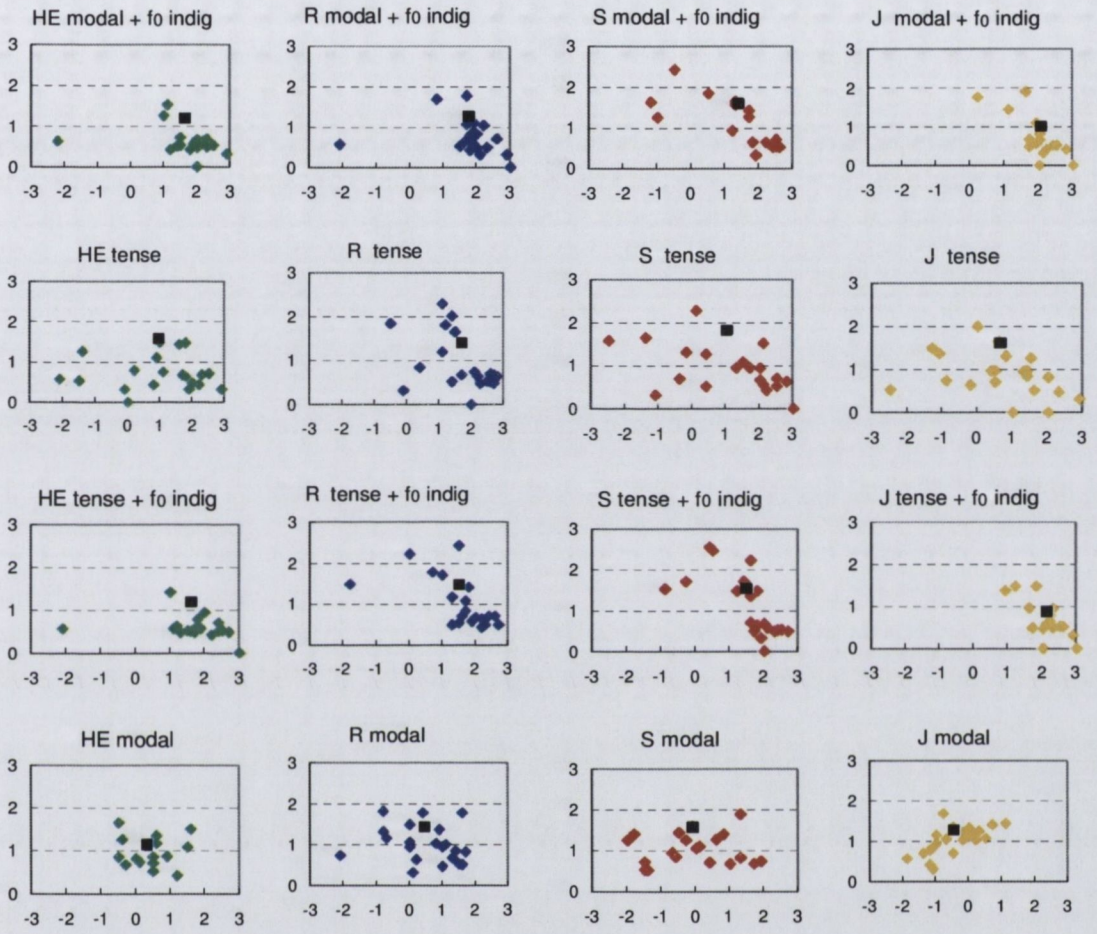


Figure 6.6. *Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for individual raters. Each dot in the plots corresponds to the data for one rater. Standard deviation of 1 and below suggests confident voice to affect association. Group mean and standard deviation are shown as a black dot.*

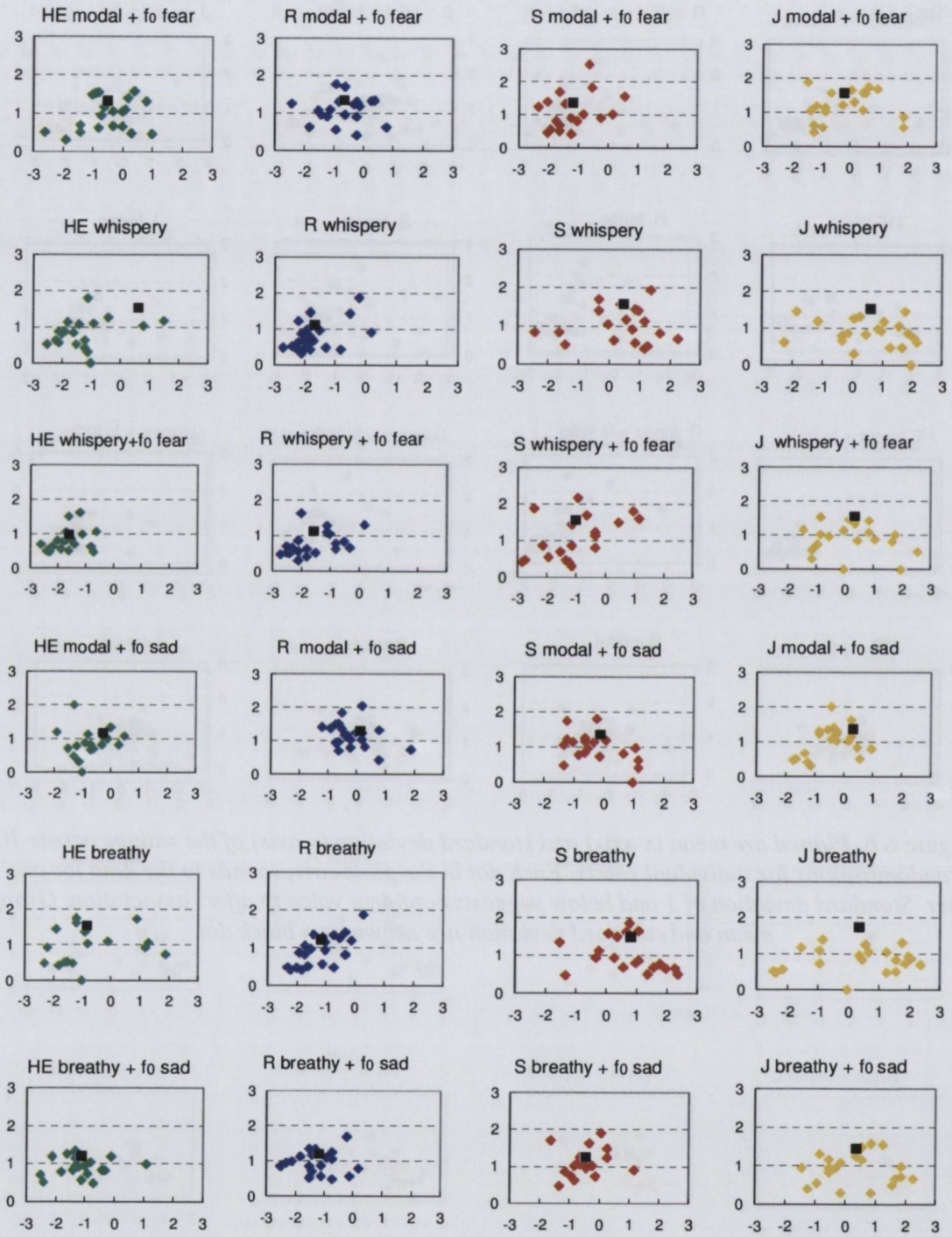


Figure 6.7. *Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for individual raters. Each dot in the plots corresponds to the data for one rater. Standard deviation of 1 and below suggests confident voice to affect association. Group mean and standard deviation are shown as a black dot.*

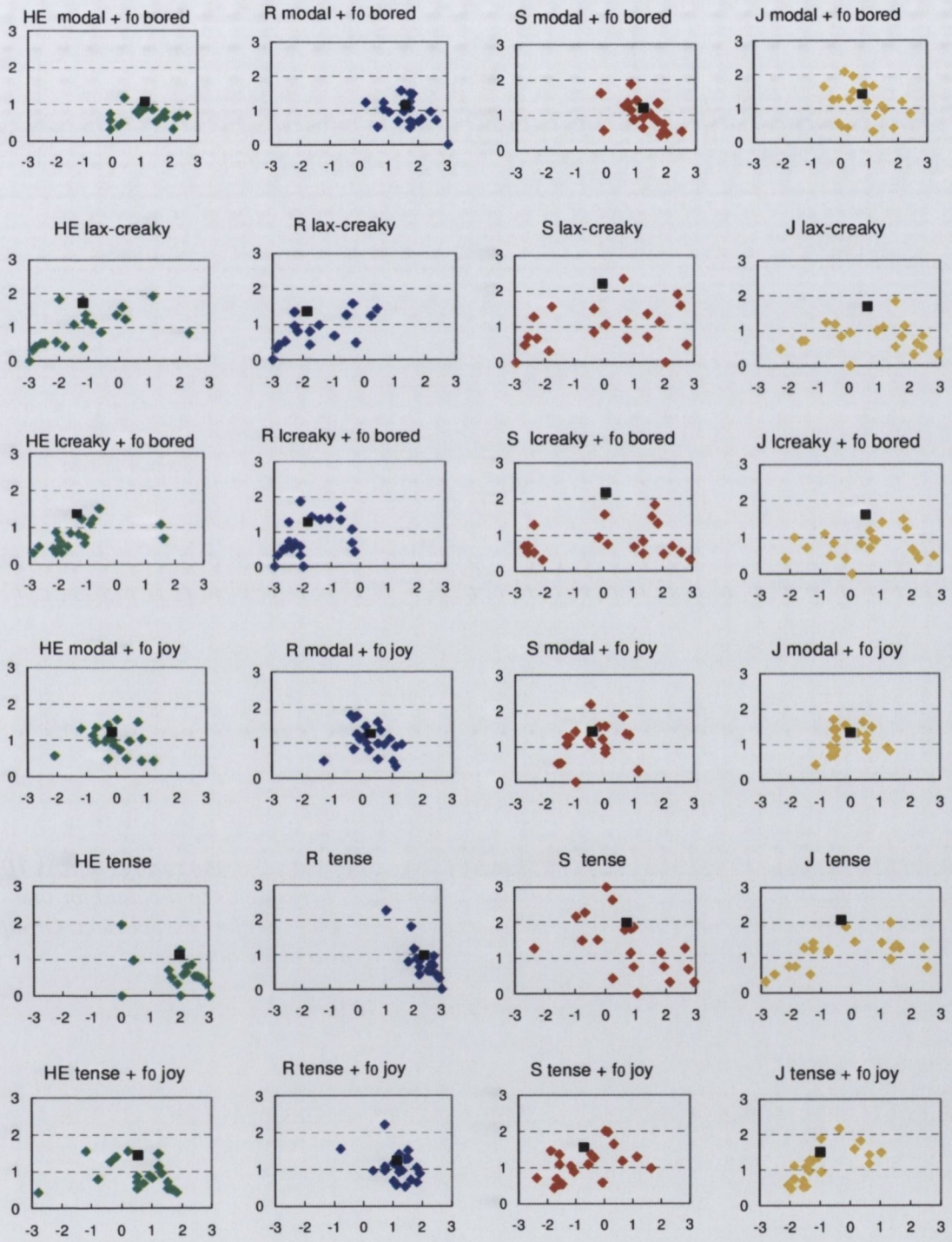


Figure 6.8. *Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for individual raters. Each dot in the plots corresponds to the data for one rater. Standard deviation of 1 and below suggests confident voice to affect association. Group mean and standard deviation are shown as a black dot.*

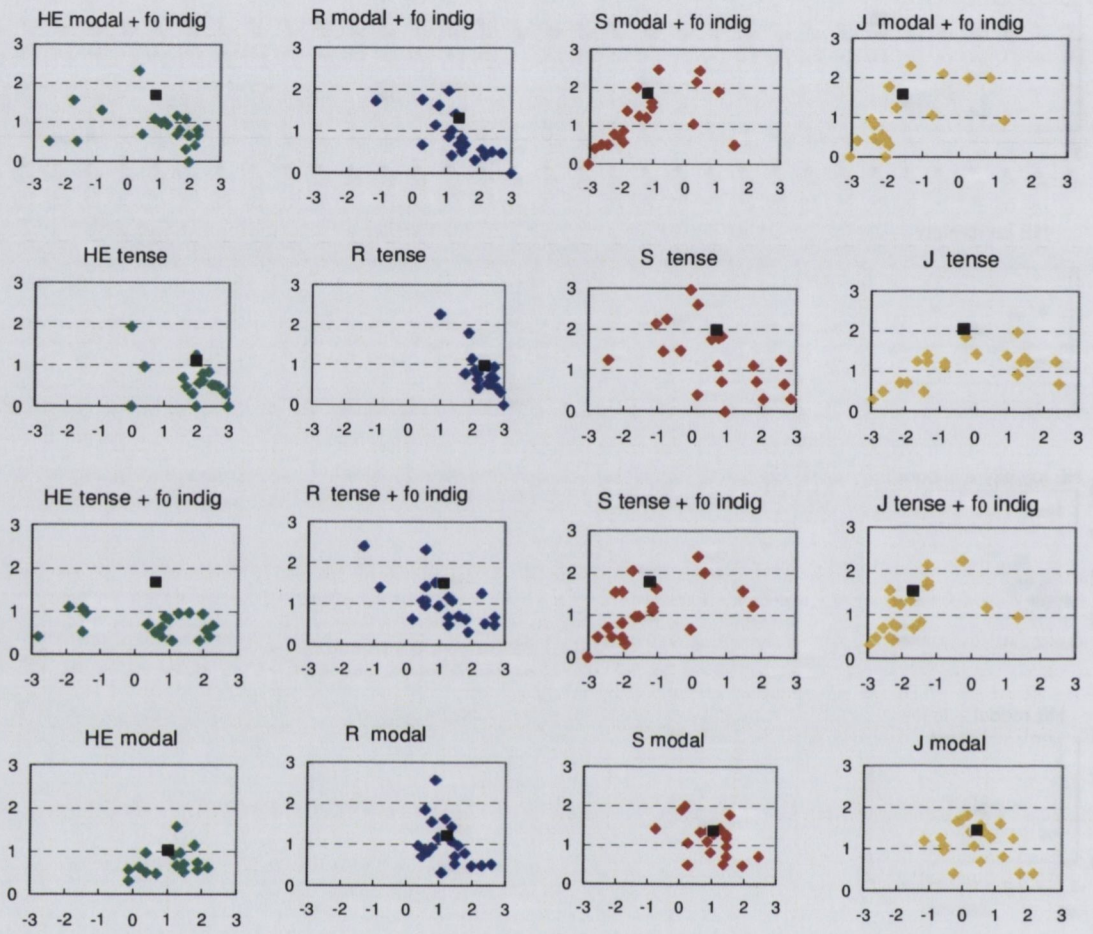


Figure 6.9. *Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for individual raters. Each dot in the plots corresponds to the data for one rater. Standard deviation of 1 and below suggests confident voice to affect association. Group mean and standard deviation are shown as a black dot.*

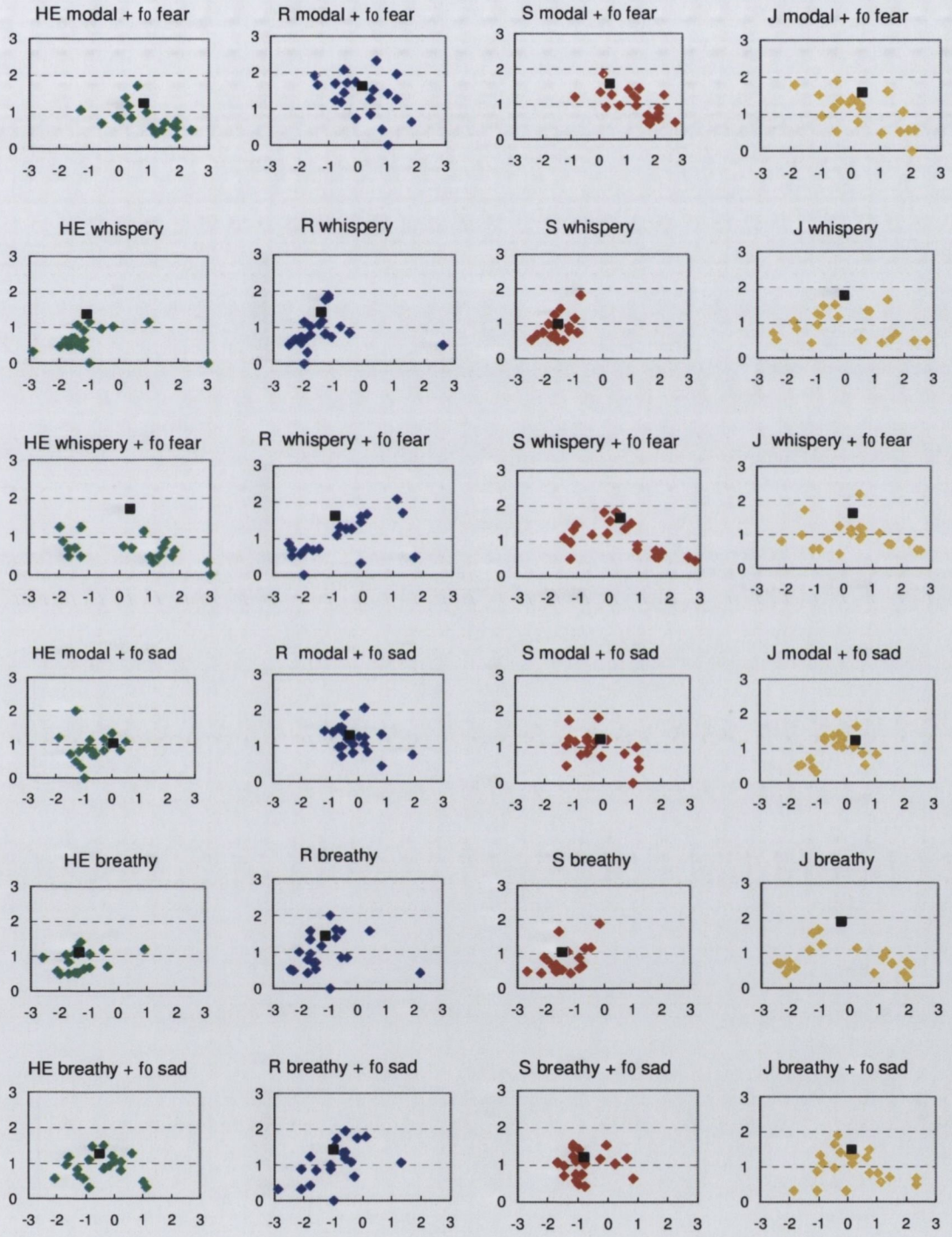


Figure 6.10. *Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for individual raters. Each dot in the plots corresponds to the data for one rater. Standard deviation of 1 and below suggests confident voice to affect association. Group mean and standard deviation are shown as a black dot.*

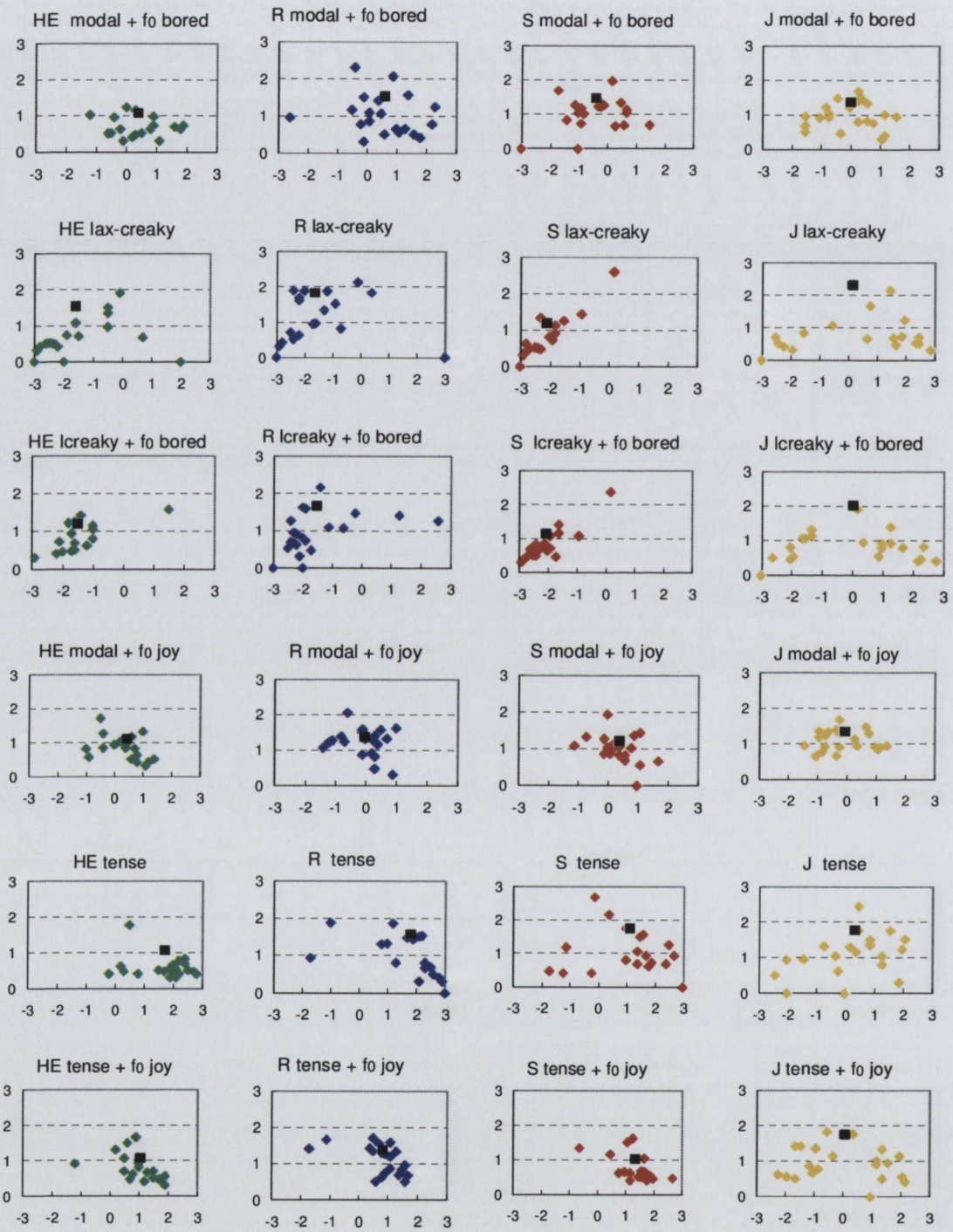


Figure 6.11. *Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for individual raters. Each dot in the plots corresponds to the data for one rater. Standard deviation of 1 and below suggests confident voice to affect association. Group mean and standard deviation are shown as a black dot.*

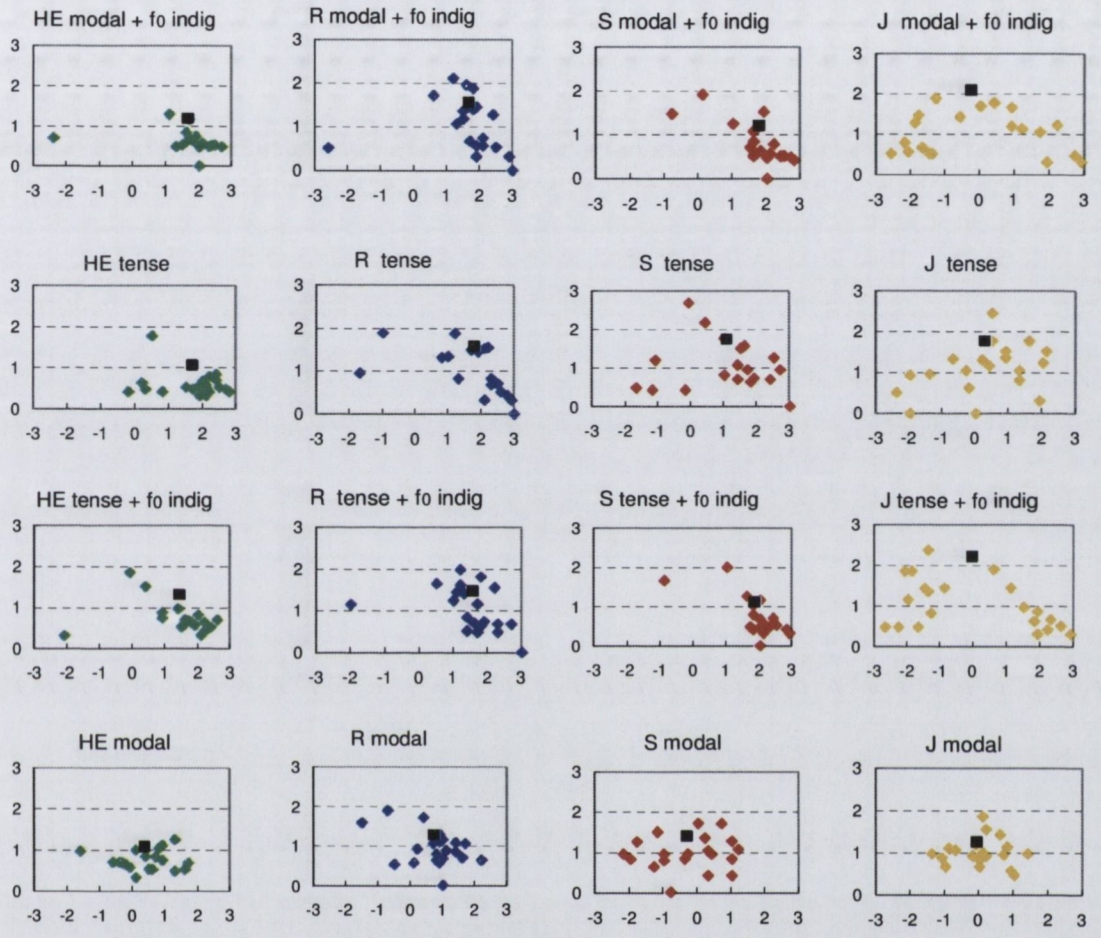


Figure 6.12. *Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for individual raters. Each dot in the plots corresponds to the data for one rater. Standard deviation of 1 and below suggests confident voice to affect association. Group mean and standard deviation are shown as a black dot.*

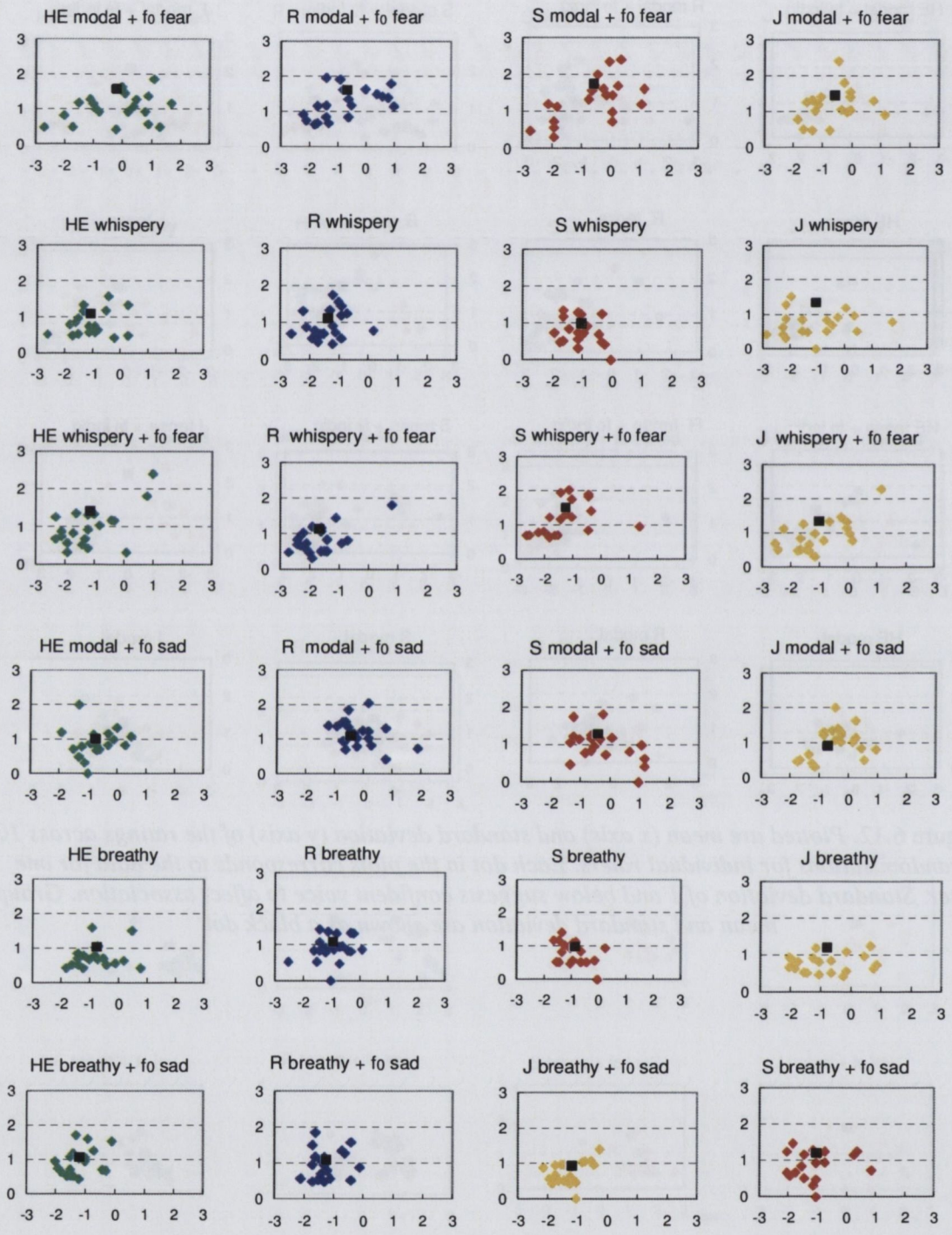


Figure 6.13. *Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for individual raters. Each dot in the plots corresponds to the data for one rater. Standard deviation of 1 and below suggests confident voice to affect association. Group mean and standard deviation are shown as a black dot.*

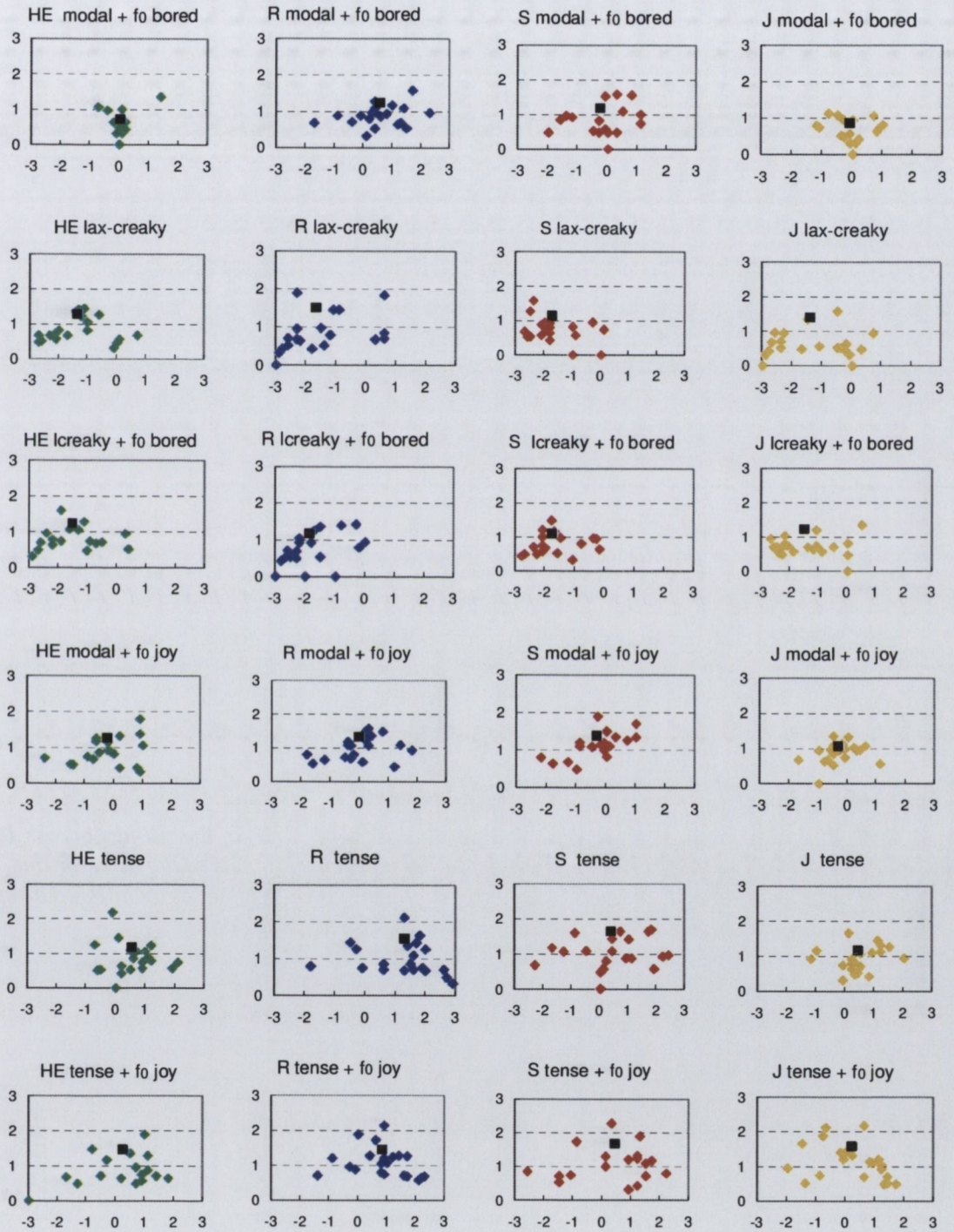


Figure 6.14. *Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for individual raters. Each dot in the plots corresponds to the data for one rater. Standard deviation of 1 and below suggests confident voice to affect association. Group mean and standard deviation are shown as a black dot.*

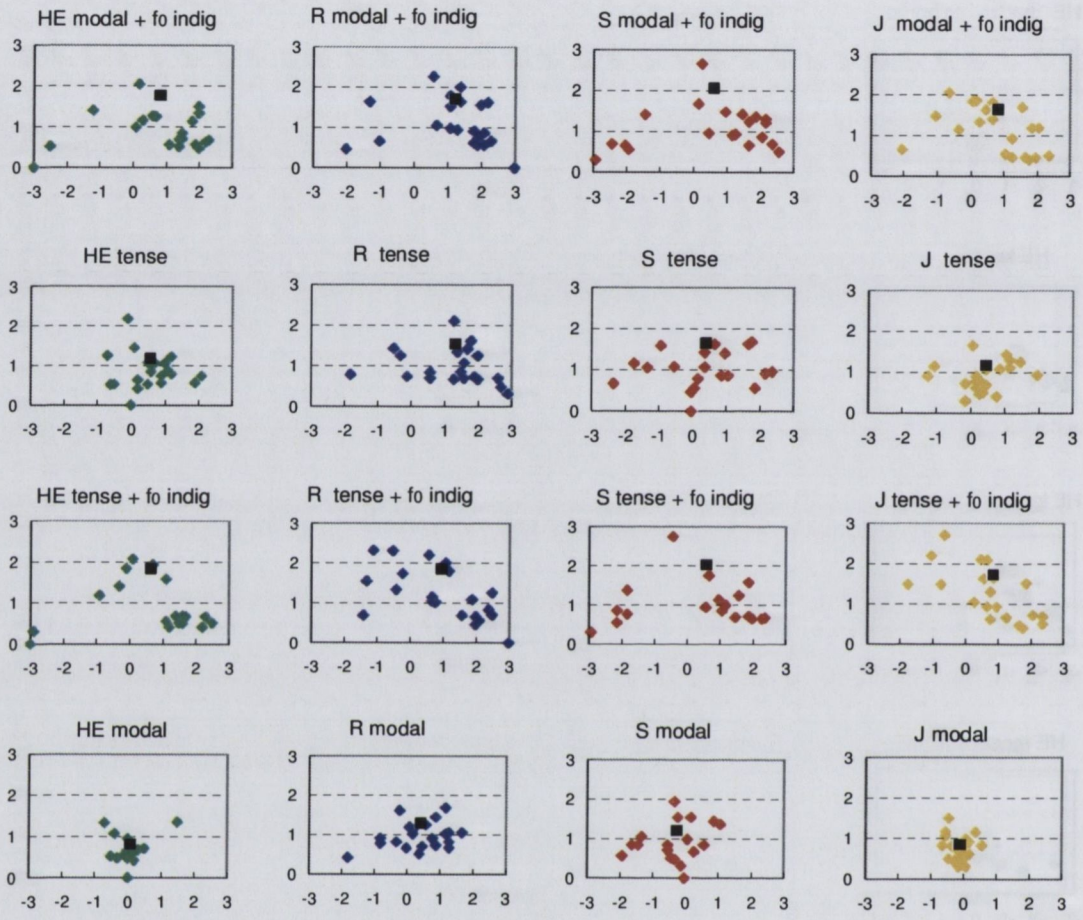


Figure 6.15. *Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for individual raters. Each dot in the plots corresponds to the data for one rater. Standard deviation of 1 and below suggests confident voice to affect association. Group mean and standard deviation are shown as a black dot.*

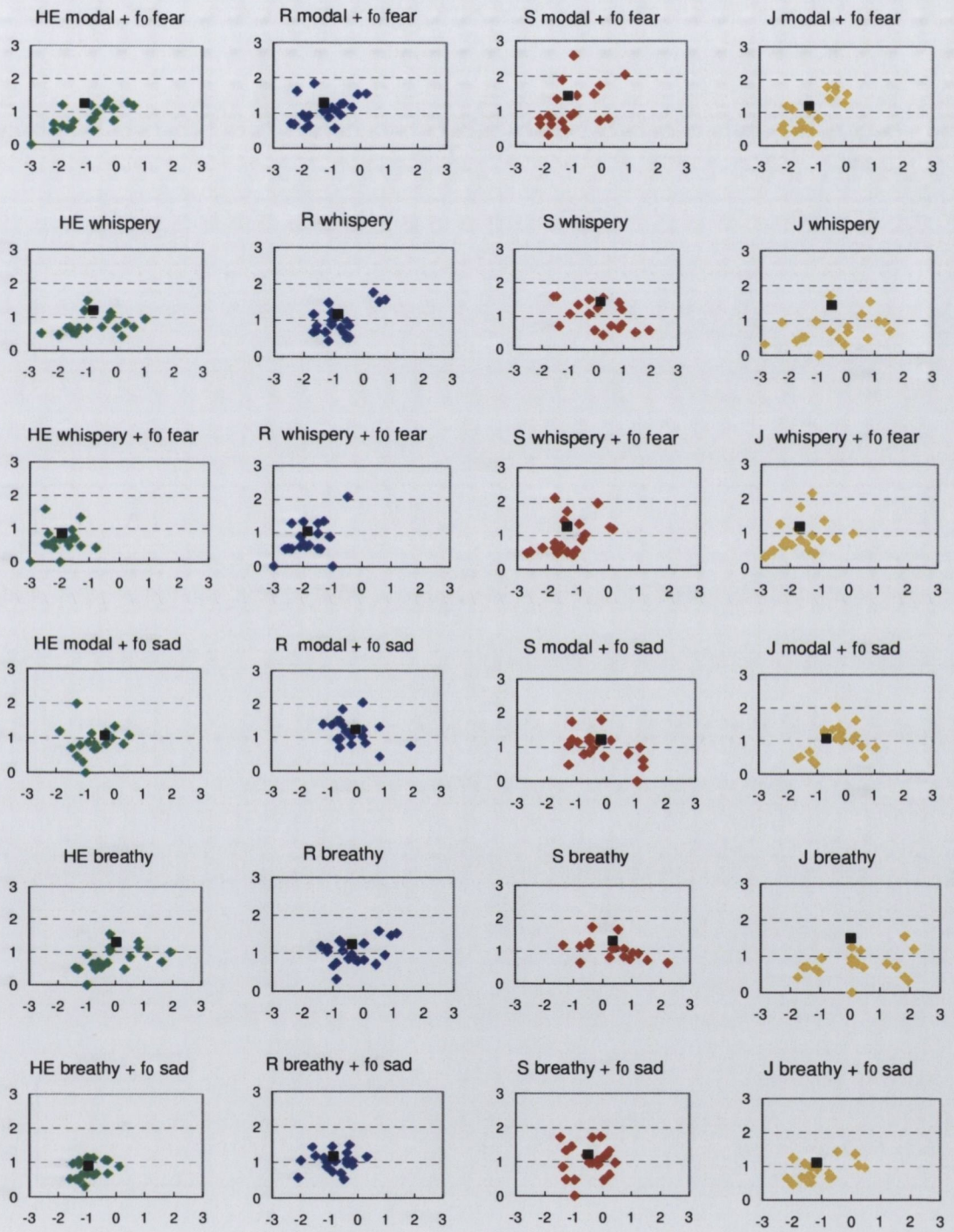


Figure 6.16. *Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for individual raters. Each dot in the plots corresponds to the data for one rater. Standard deviation of 1 and below suggests confident voice to affect association. Group mean and standard deviation are shown as a black dot.*

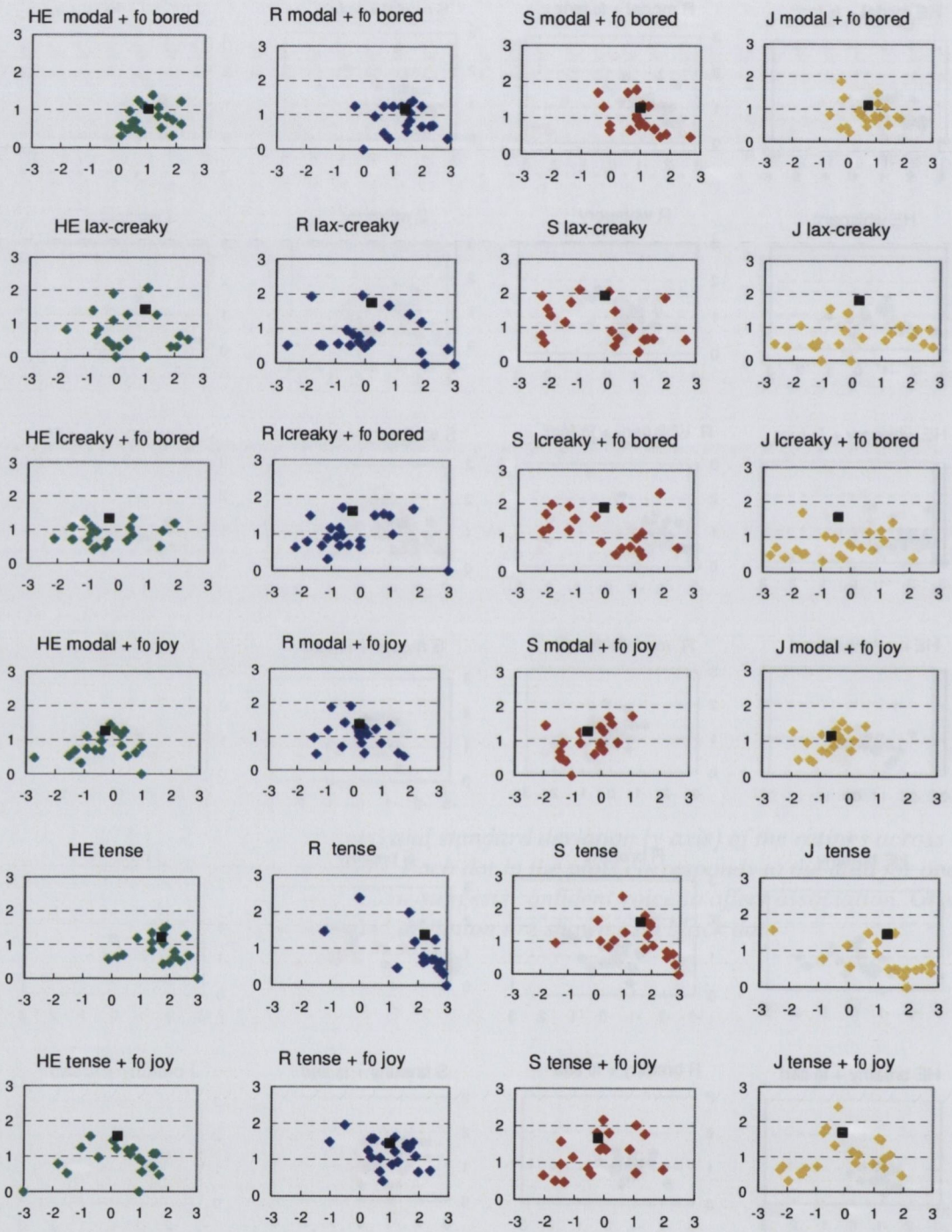


Figure 6.17. *Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for individual raters. Each dot in the plots corresponds to the data for one rater. Standard deviation of 1 and below suggests confident voice to affect association. Group mean and standard deviation are shown as a black dot.*

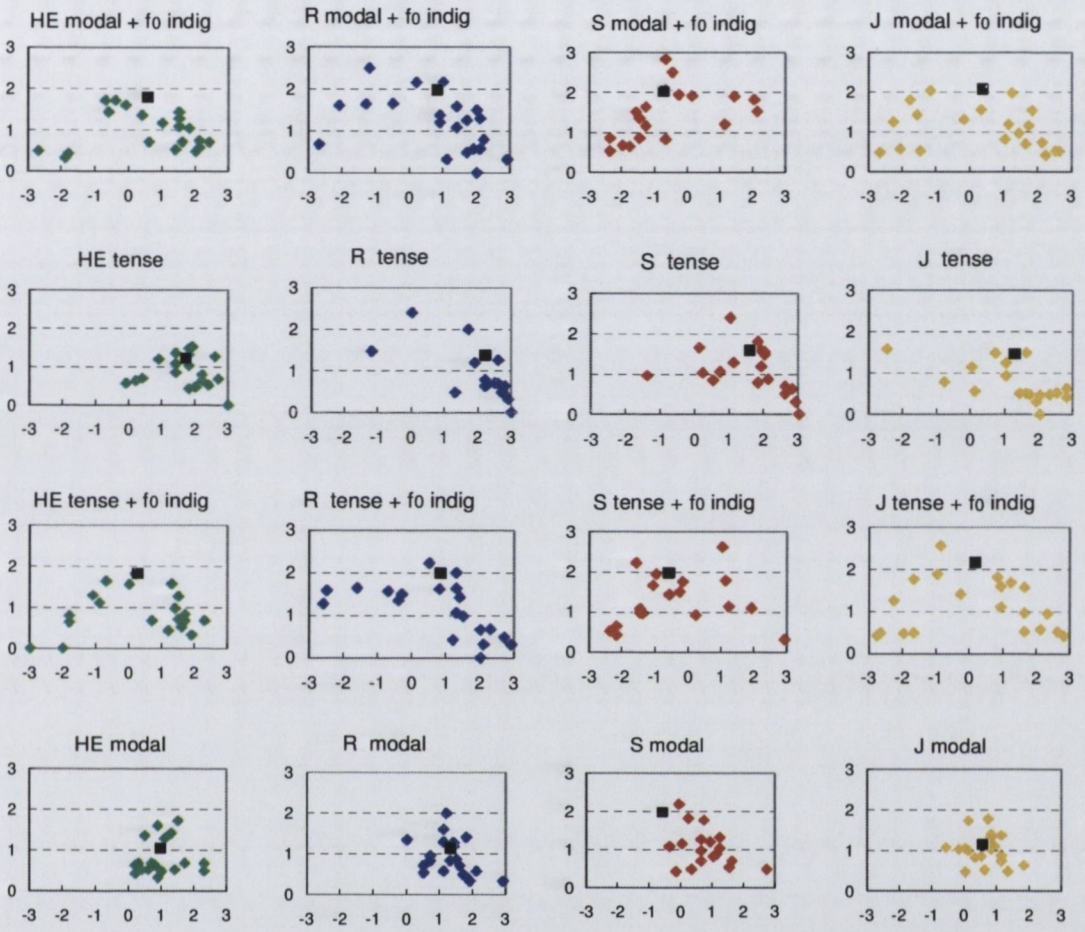


Figure 6.18. *Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for individual raters. Each dot in the plots corresponds to the data for one rater. Standard deviation of 1 and below suggests confident voice to affect association. Group mean and standard deviation are shown as a black dot.*

7 Appendix 7

Summary of results for the cross-language study (Chapter 4).

Figures 7.1 – 7.12 expand on Figures 4.7 and 4.8 in Chapter 4 by including for each of the subtests representations that allow us to look at not only **cross-language agreement** (shown in red boxes) but also: cases where **three languages** show above threshold responses for particular stimuli (blue boxes); cases where **two languages** show above threshold responses for particular stimuli (grey boxes); or even **a single language** (orange boxes). These latter figures are useful in allowing us to focus on the prevalence of cross-language differences.

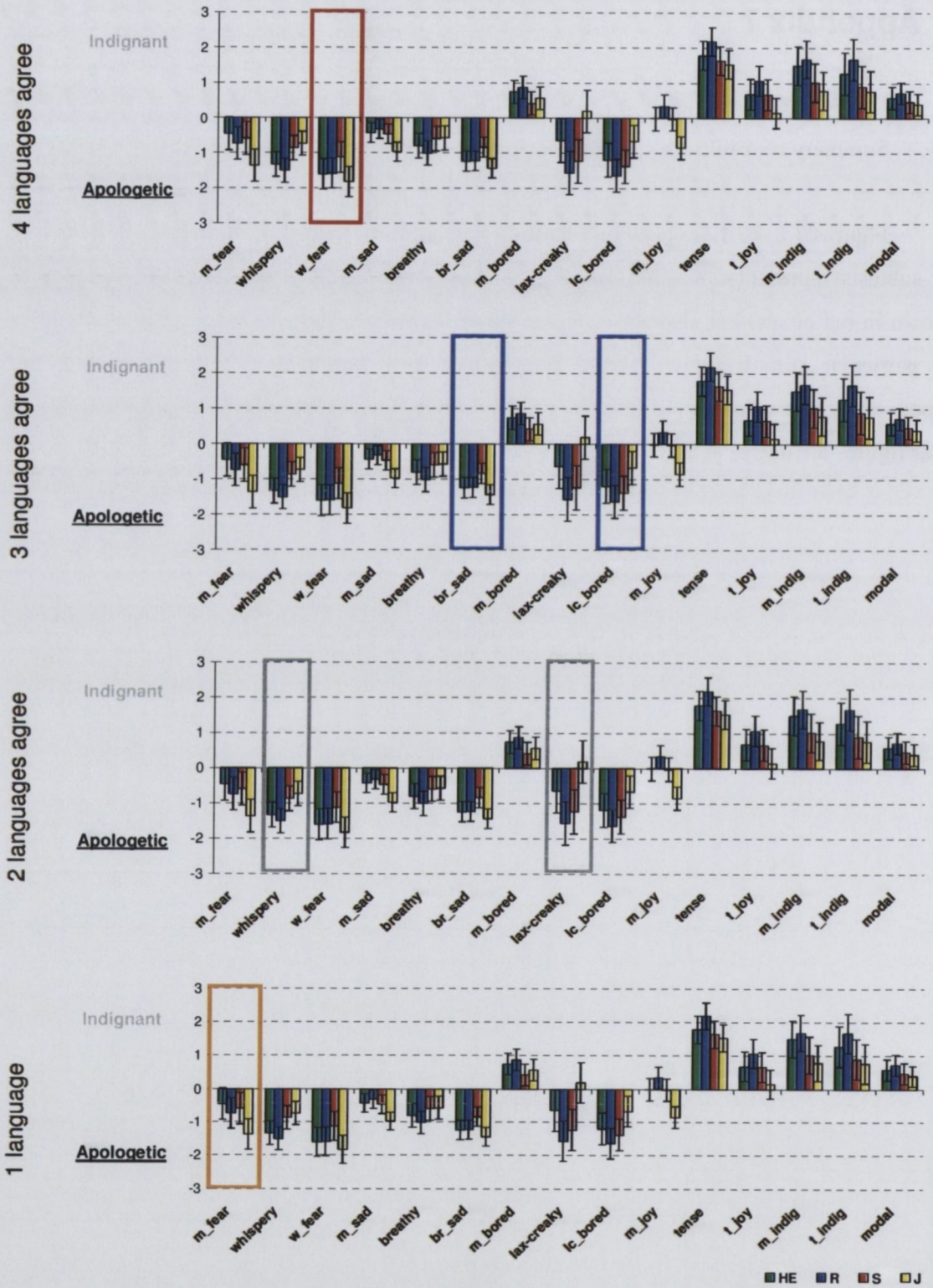


Figure 7.1. *Language convergence/divergence, for the* apologetic *affect.*

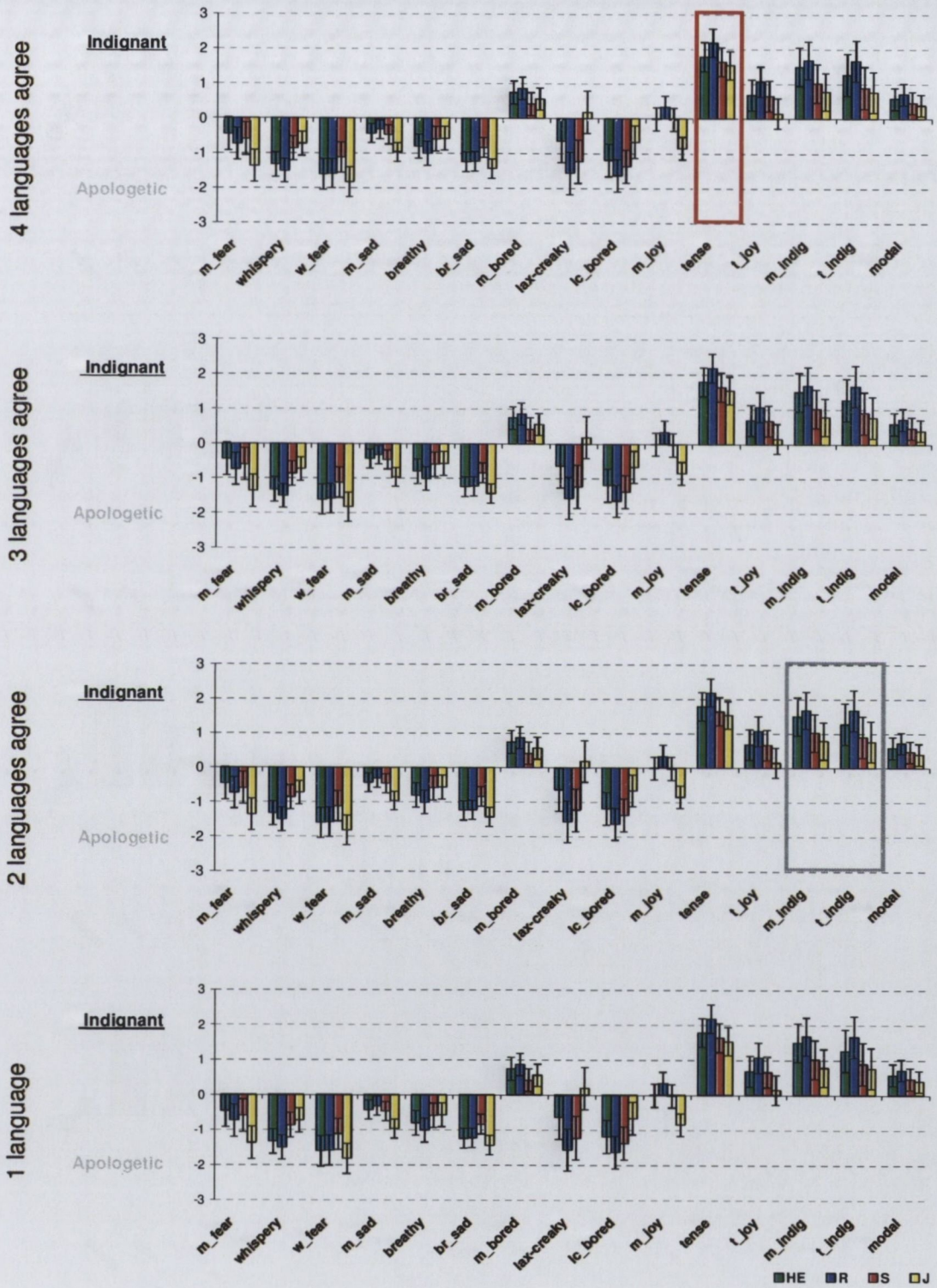


Figure 7.2. *Language convergence/divergence, for the* indignant *affect.*

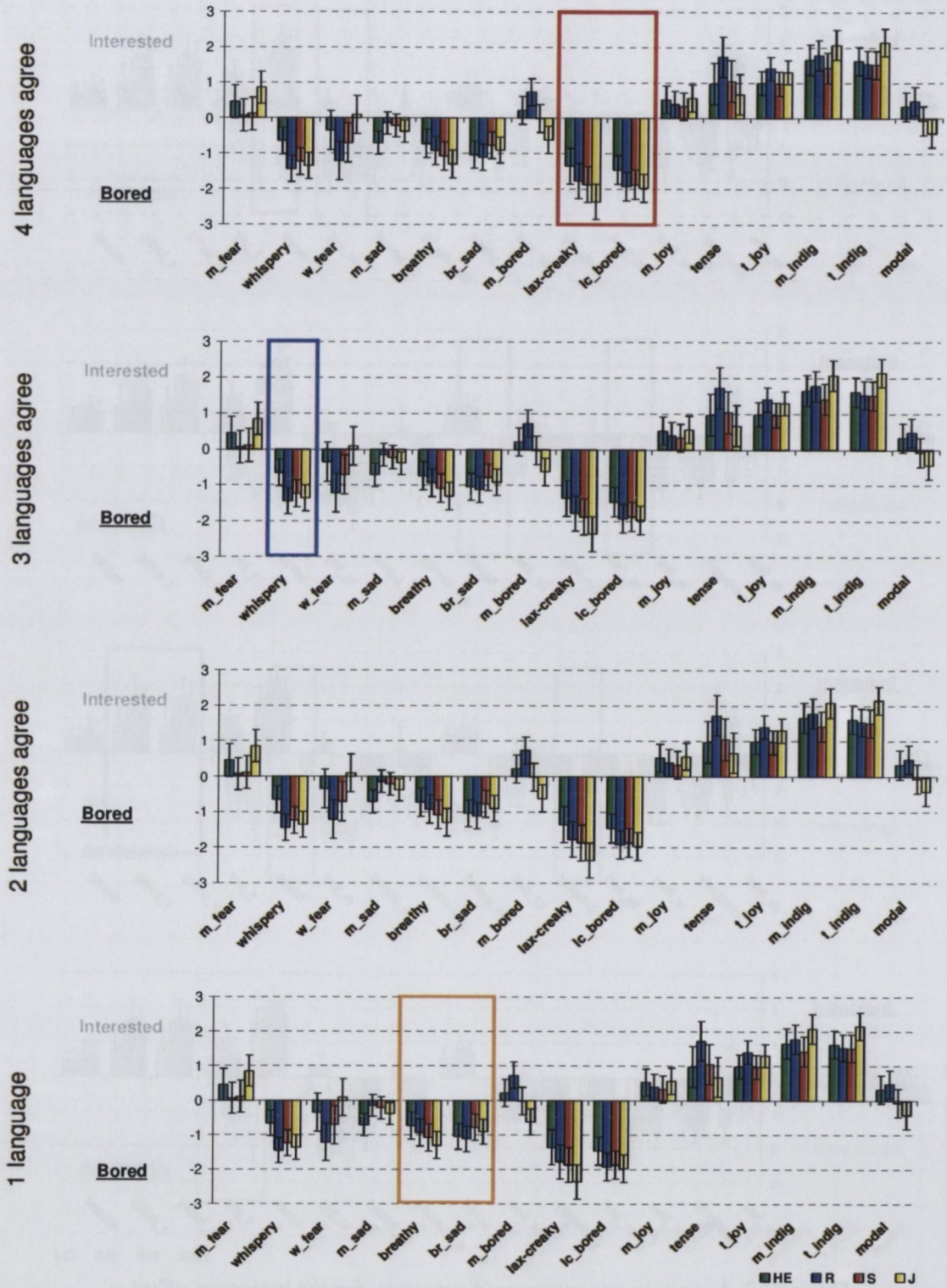


Figure 7.3. *Language convergence/divergence, for the* bored *affect.*

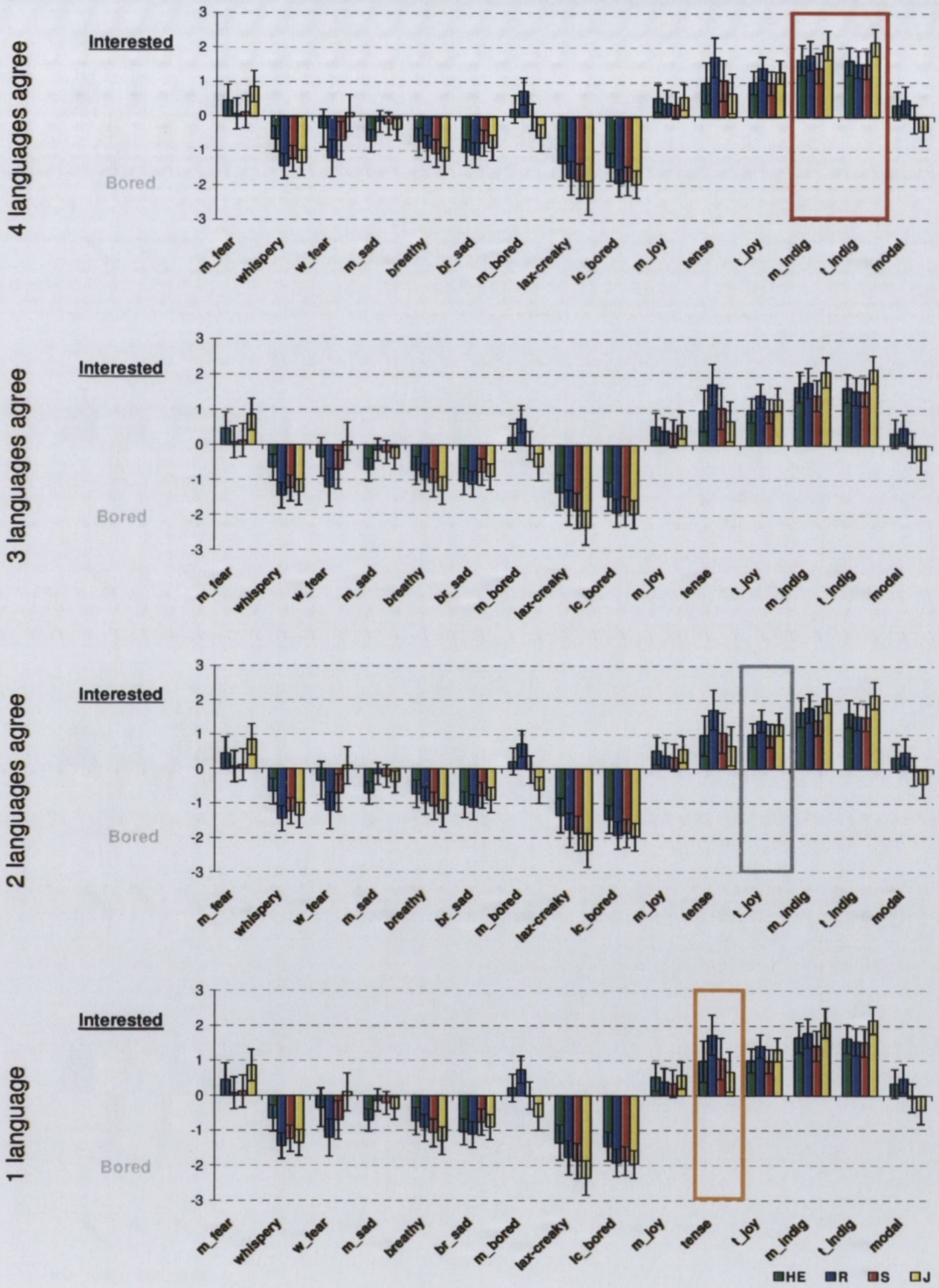


Figure 7.4. *Language convergence/divergence, for the* interested *affect.*

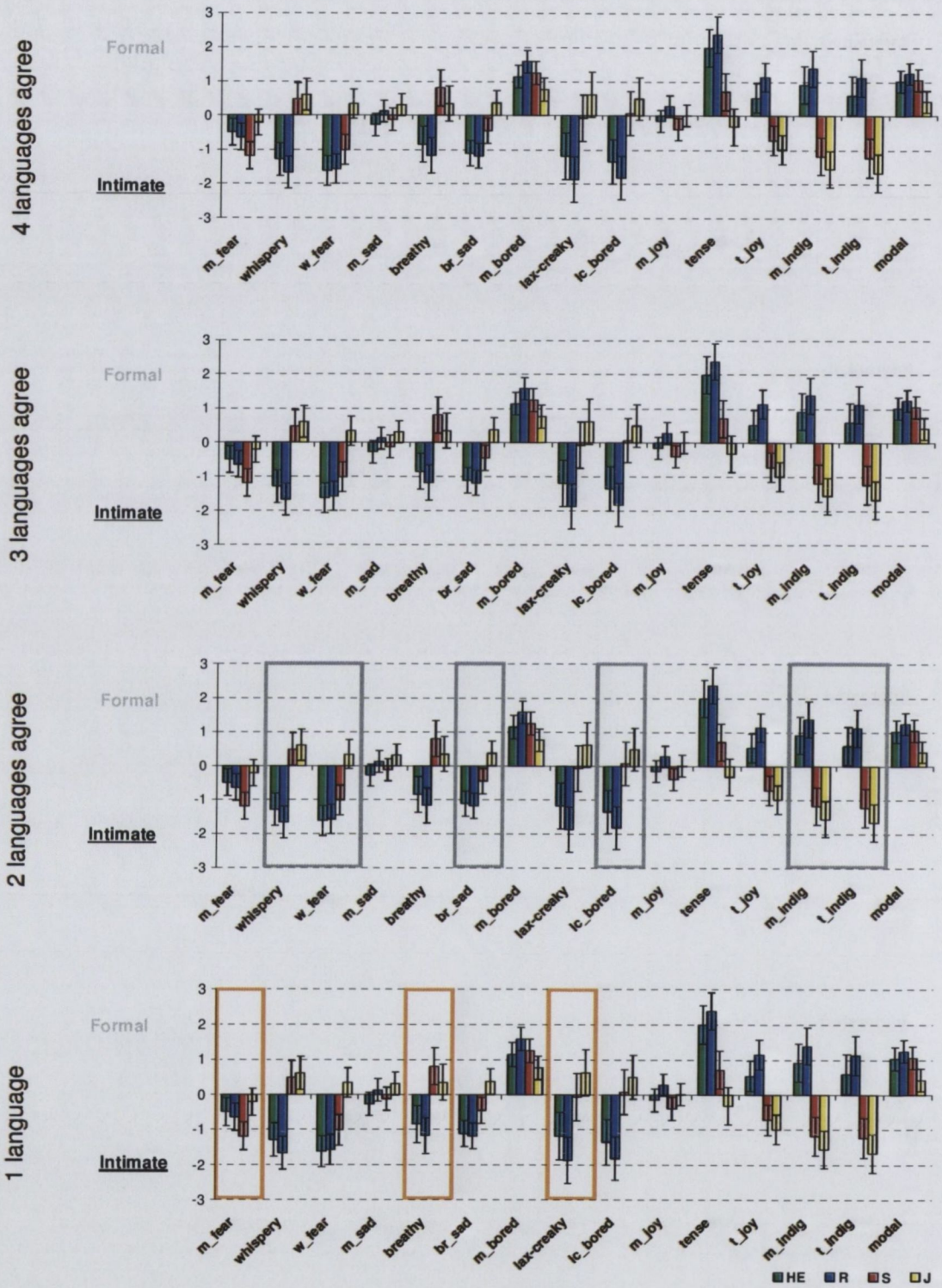


Figure 7.5. *Language convergence/divergence, for the* intimate *affect.*

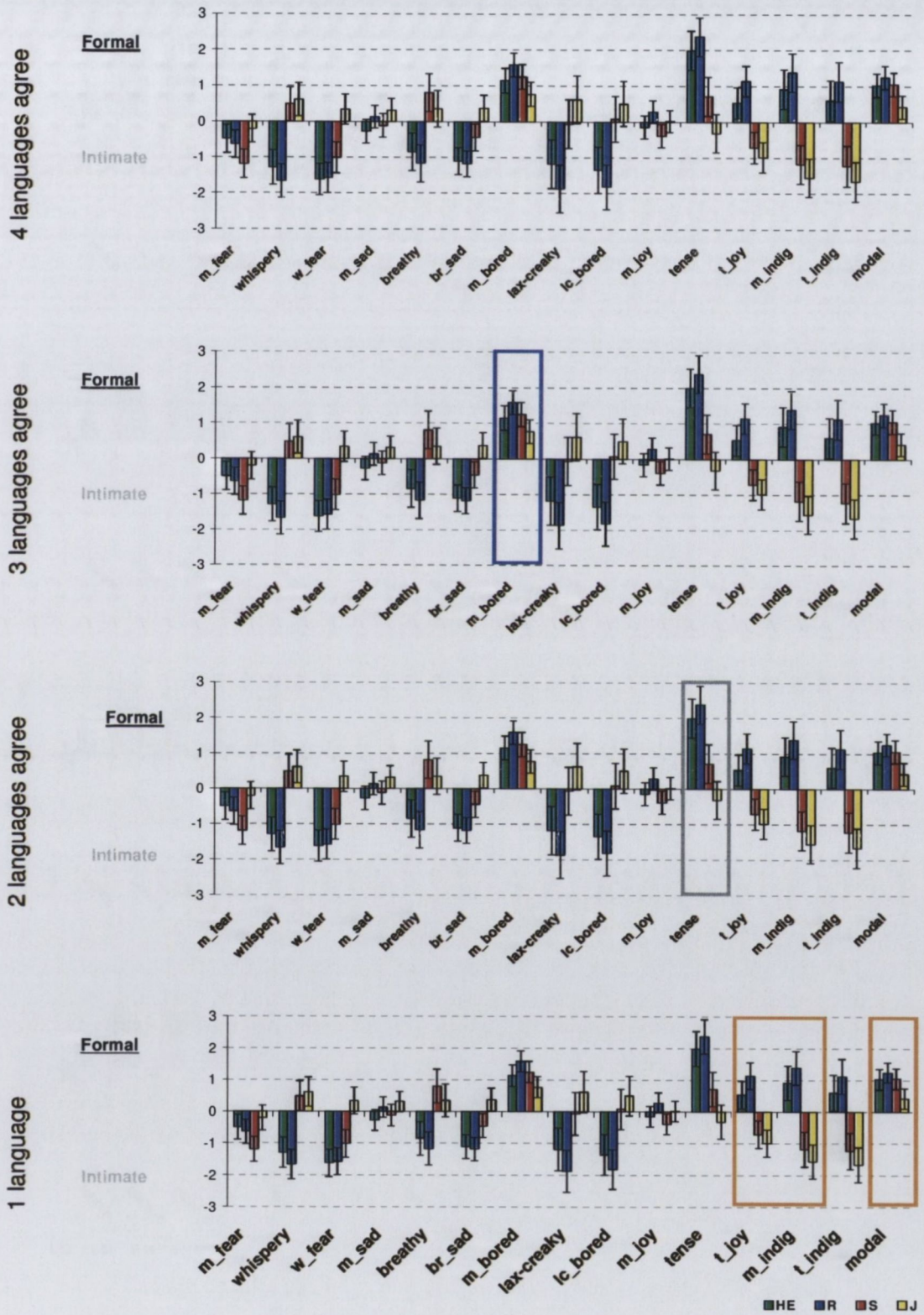


Figure 7.6. *Language convergence/divergence, for the* formal *affect.*

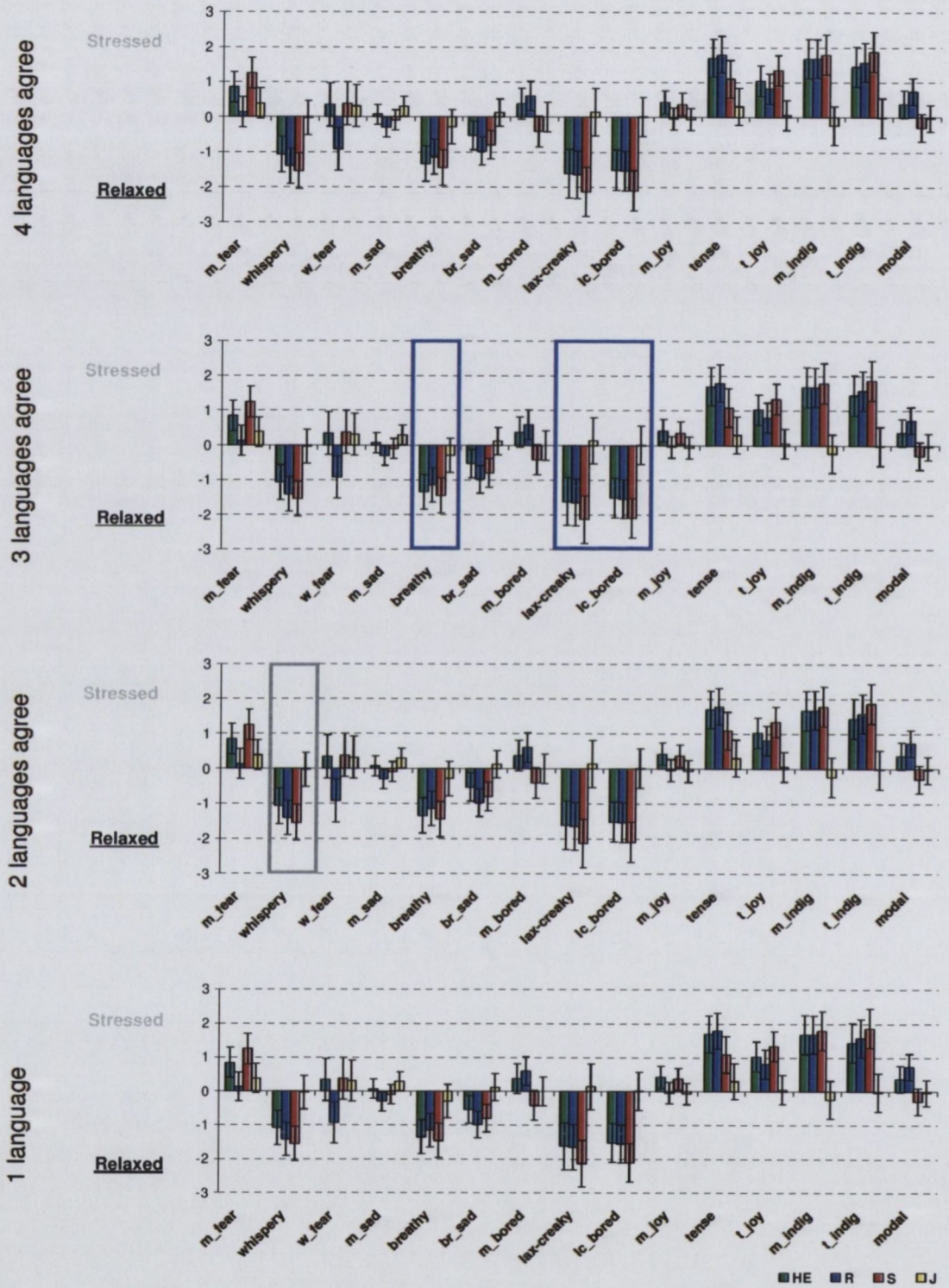


Figure 7.7. *Language convergence/divergence, for the* relaxed *affect.*

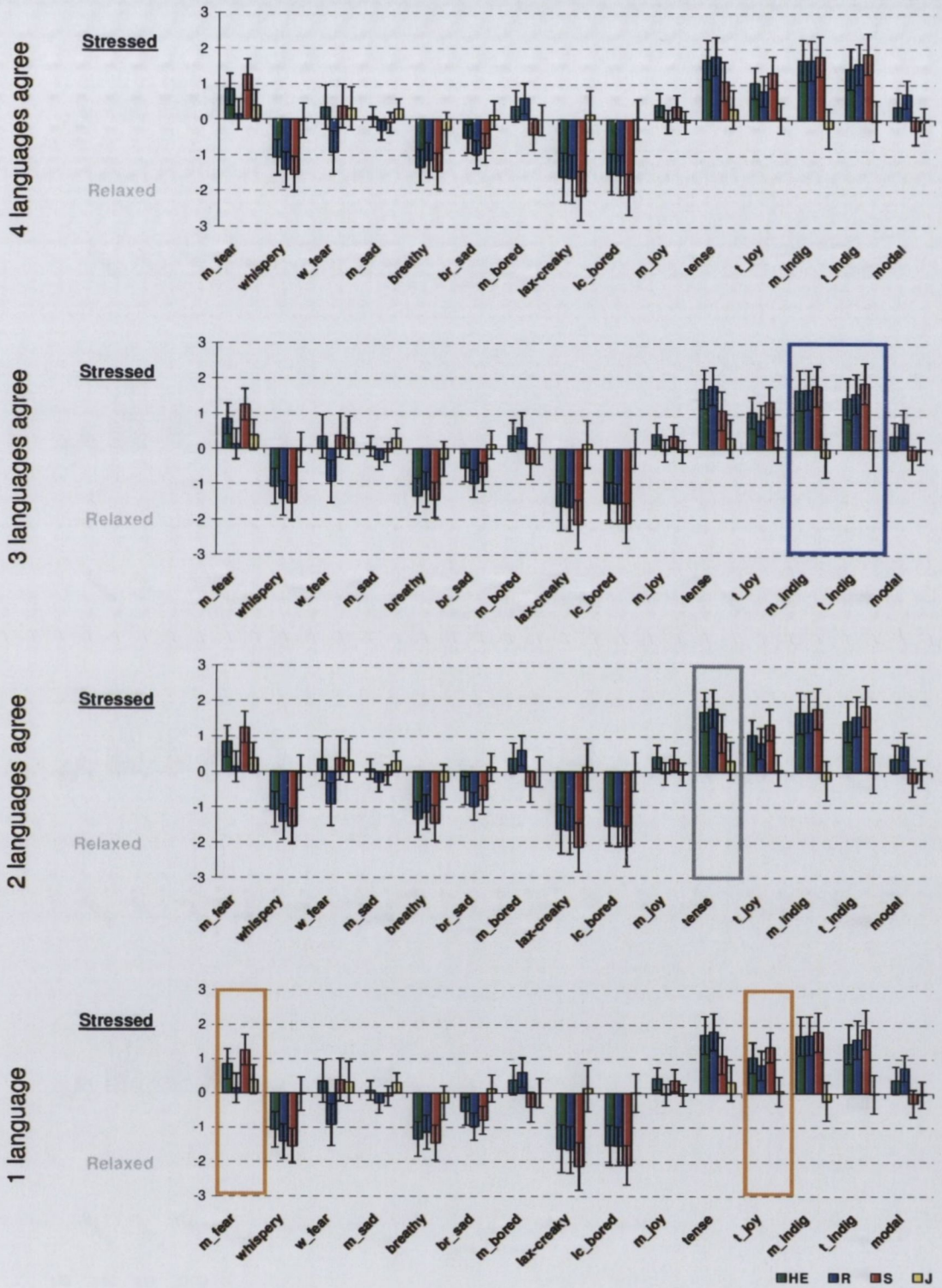


Figure 7.8. *Language convergence/divergence, for the* stressed *affect.*

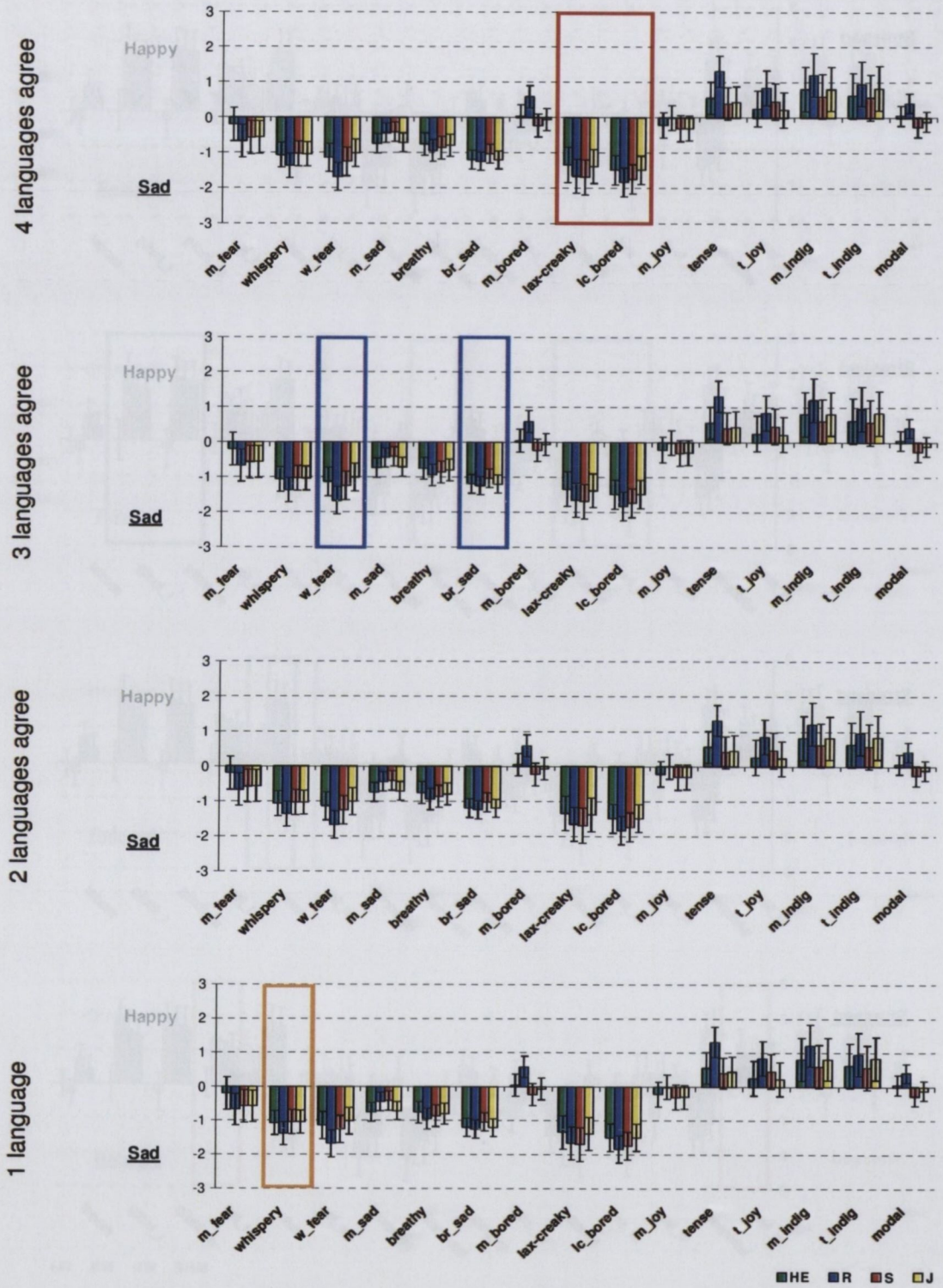


Figure 7.9. Language convergence/divergence, for the sad *affect.*

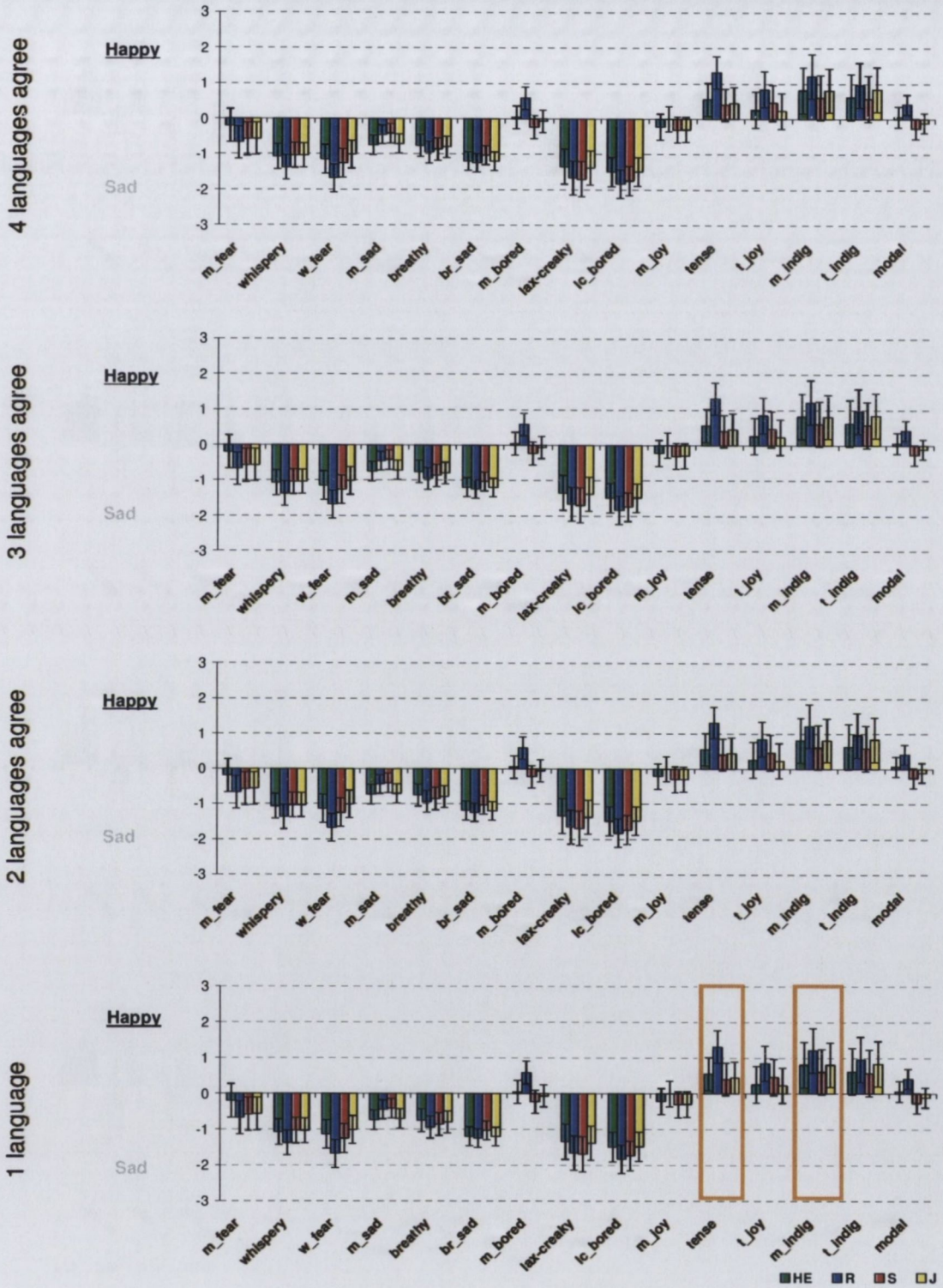


Figure 7.10. *Language convergence/divergence, for the* happy *affect.*

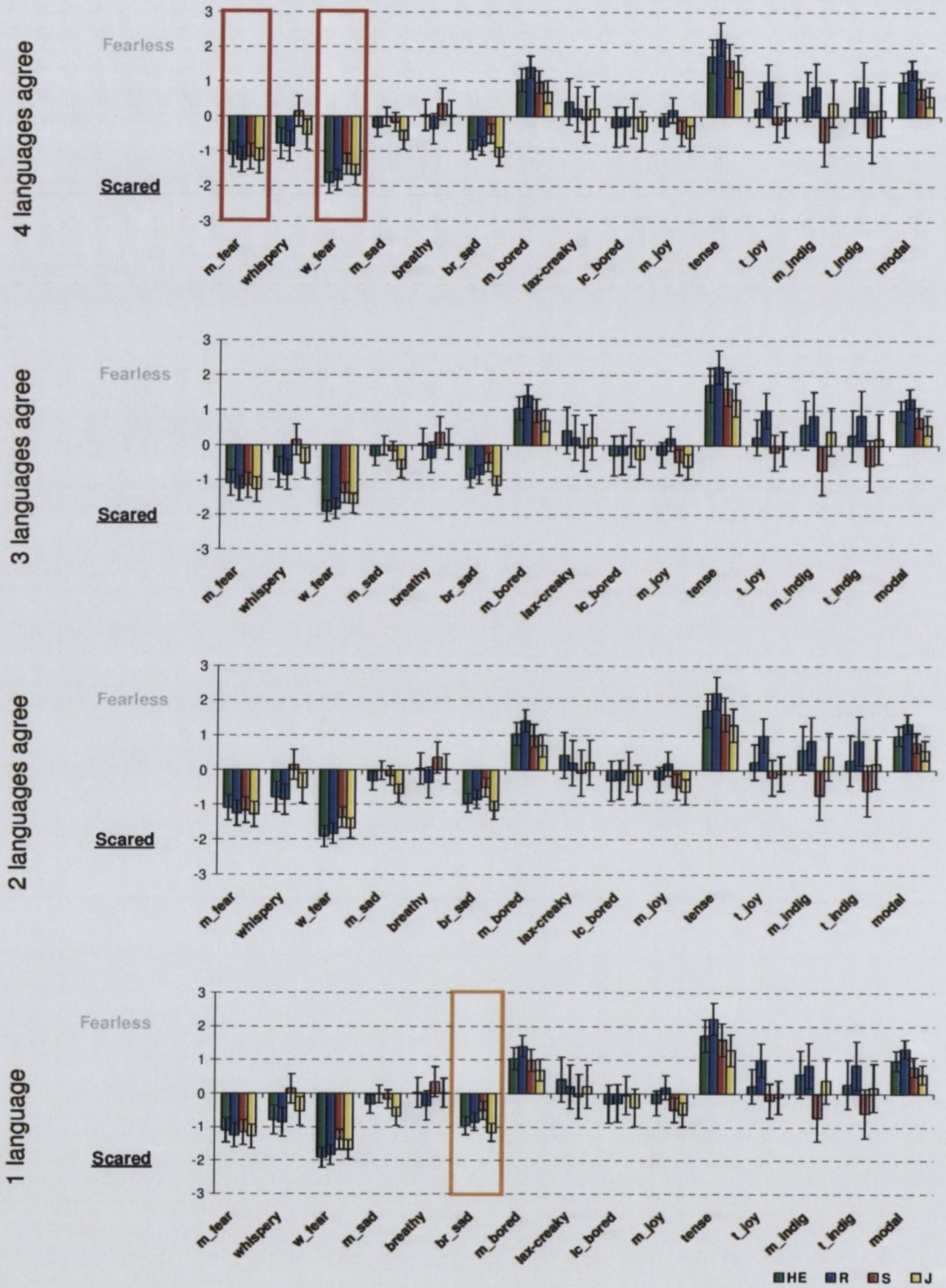


Figure 7.11. *Language convergence/divergence, for the* scared *affect.*

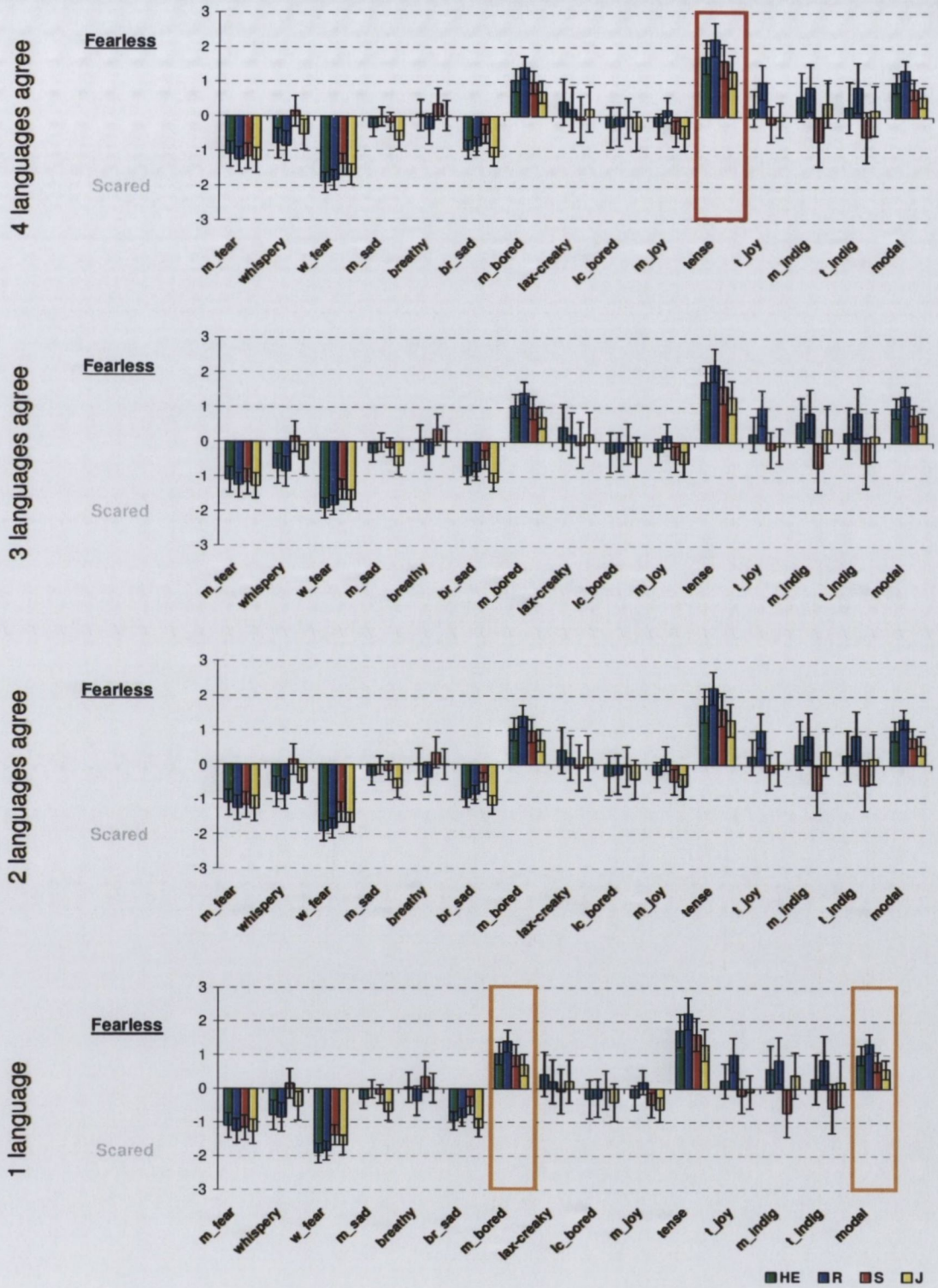


Figure 7.12. *Language convergence/divergence, for the* fearless *affect.*

8 Appendix 8

Ratings obtained in the Loudness experiment 1 (Chapter 5).

For each stimulus, in the box plots shown are mean (dot), median (line), interquartile range (box) and extreme values (whiskers).

Red rectangles highlight the cases where affective ratings yielded by the stimuli were above 1, corresponding to mild to strong affective colouring.

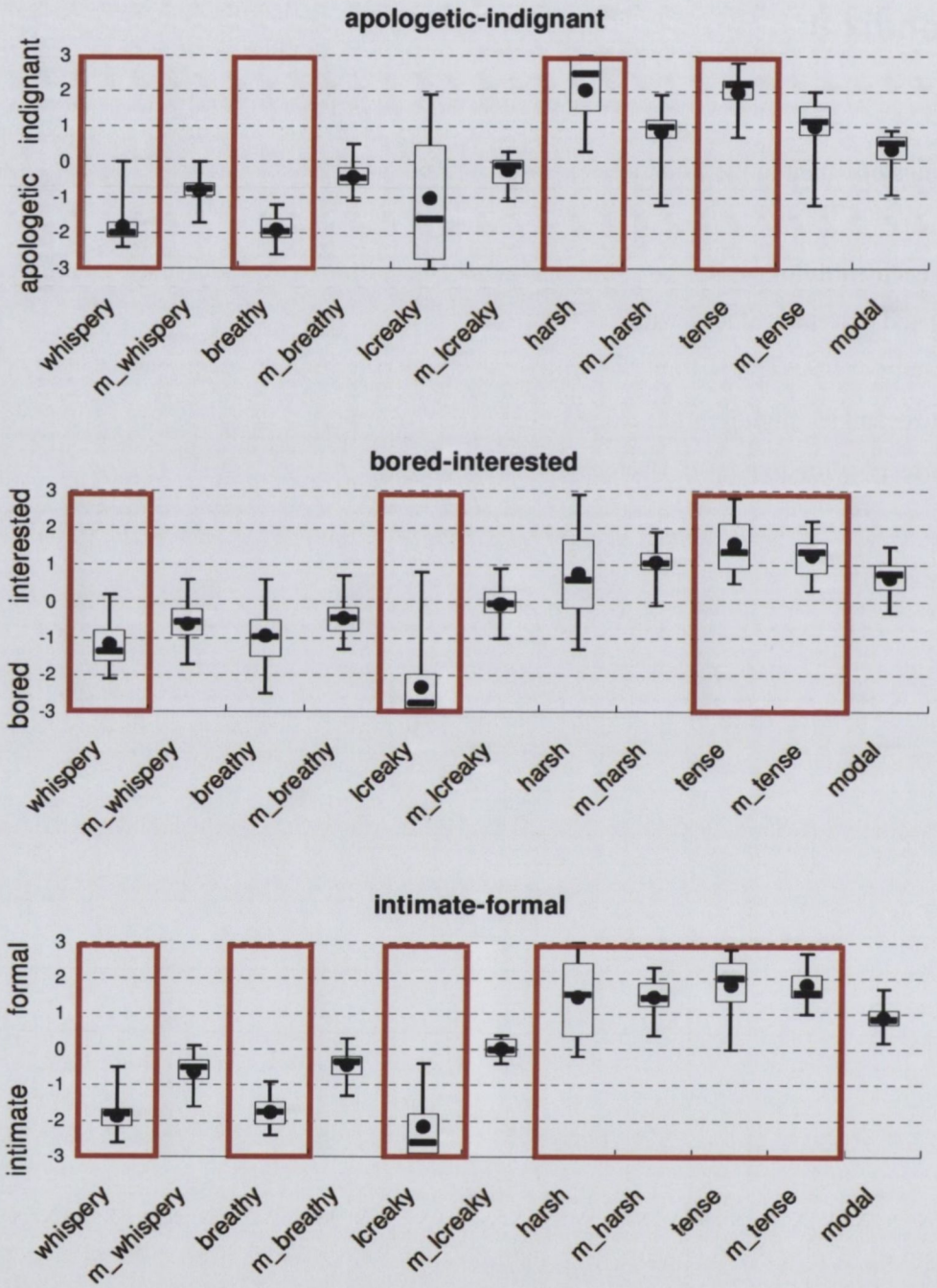


Figure 8.1. *Responses obtained for the voice quality stimuli and loudness stimuli for each subtest. Loudness 1. The plots show interquartile range (box), mean (filled circle), median (horizontal line in box) and extreme values (whiskers).*

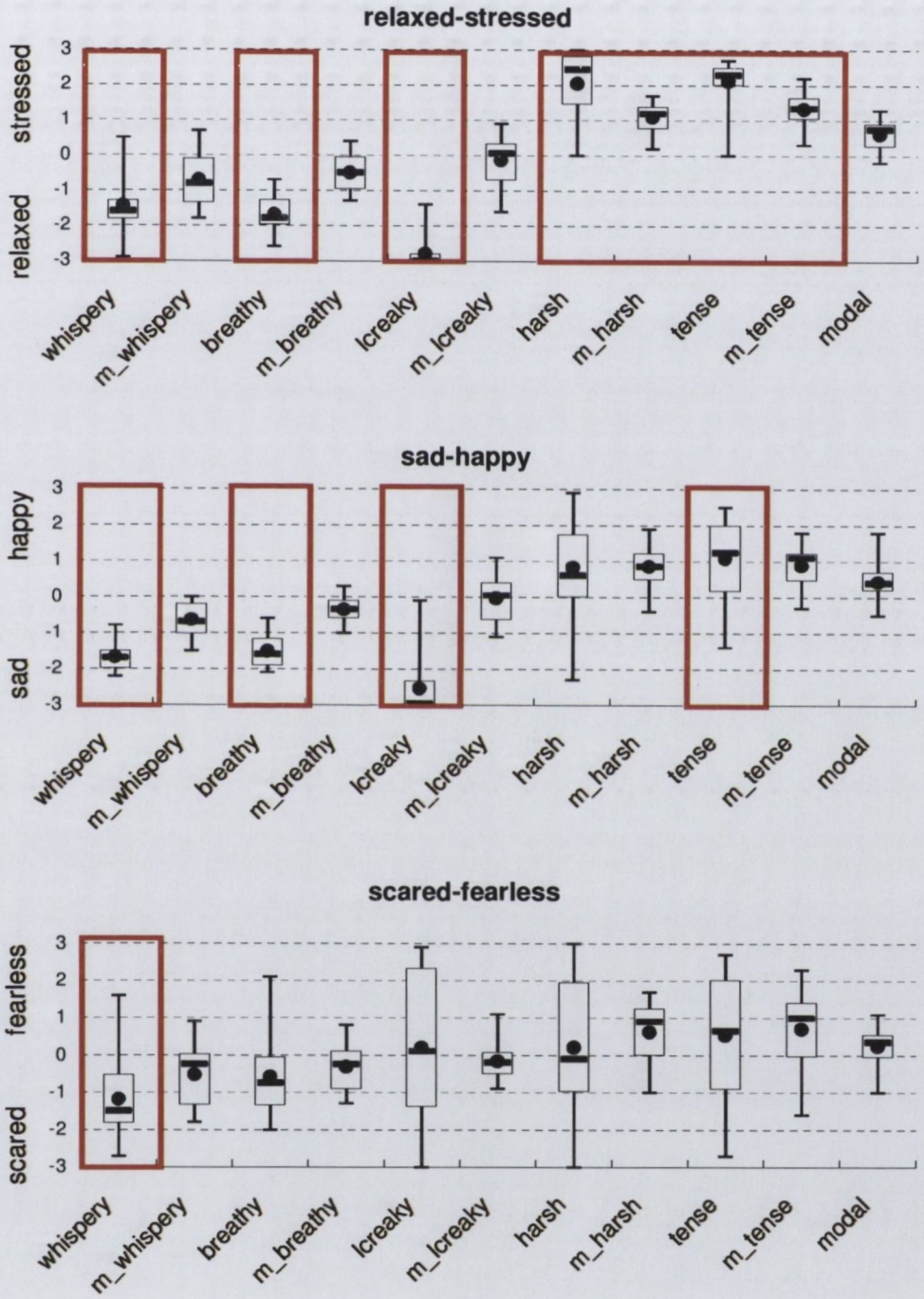


Figure 8.2. *Responses obtained for the voice quality stimuli and loudness stimuli for each subtest (continued). Loudness 1. The plots show interquartile range (box), mean (filled circle), median (horizontal line in box) and extreme values (whiskers).*

9 Appendix 9

Loudness experiment 1 ANOVA

Apologetic-Indignant

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
loudness	Sphericity Assumed	249.081	4	62.270	80.624	.000	.843
	Greenhouse-Geisser	249.081	2.080	119.757	80.624	.000	.843
Error(loudness)	Sphericity Assumed	46.341	60	.772			
	Greenhouse-Geisser	46.341	31.198	1.485			
voice	Sphericity Assumed	2.139	1	2.139	1.591	.226	.096
	Greenhouse-Geisser	2.139	1.000	2.139	1.591	.226	.096
Error(voice)	Sphericity Assumed	20.168	15	1.345			
	Greenhouse-Geisser	20.168	15.000	1.345			
loudness*voice	Sphericity Assumed	47.479	4	11.870	23.109	.000	.606
	Greenhouse-Geisser	47.479	1.849	25.680	23.109	.000	.606
Error(loudness* voice)	Sphericity Assumed	30.819	60	.514			
	Greenhouse-Geisser	30.819	27.733	1.111			

Apologetic-Indignant: Pairwise Comparisons I

loudness	(I) voice	(J) voice	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
						Lower Bound	Upper Bound
1	whispery	m_whispery	-1.019*	.173	.000	-1.389	-.649
2	breathy	m_breathy	-1.488*	.151	.000	-1.810	-1.165
3	lax-creaky	m_lax-creaky	-.794	.500	.133	-1.858	.271
4	harsh	m_harsh	1.181*	.297	.001	.548	1.814
5	tense	m_tense	.962*	.185	.000	.568	1.357

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

a. Adjustment for multiple comparisons: Bonferroni.

Apologetic-Indignant: Pairwise Comparisons II

voice	(I) loudness	(J) loudness	Mean Difference (I-J)	Std. Error	Sig.[a]	95% C for Difference[a]	
						Lower Bound	Upper Bound
original	whispery	breathy	.088	.166	1.000	-.457	.632
		lax-creaky	-.806	.466	1.000	-2.339	.726
		harsh	-3.862*	.338	.000	-4.972	-2.753
		tense	-3.794*	.276	.000	-4.700	-2.887
	breathy	whispery	-.088	.166	1.000	-.632	.457
		lax-creaky	-.894	.510	.999	-2.569	.781
		harsh	-3.950*	.260	.000	-4.805	-3.095
		tense	-3.881*	.160	.000	-4.408	-3.355
	lax-creaky	whispery	.806	.466	1.000	-.726	2.339
		breathy	.894	.510	.999	-.781	2.569
		harsh	-3.056*	.493	.000	-4.675	-1.437
		tense	-2.988*	.511	.000	-4.667	-1.308
	harsh	whispery	3.862*	.338	.000	2.753	4.972
		breathy	3.950*	.260	.000	3.095	4.805
		lax-creaky	3.056*	.493	.000	1.437	4.675
		tense	.069	.169	1.000	-.488	.626
	tense	whispery	3.794*	.276	.000	2.887	4.700
		breathy	3.881*	.160	.000	3.355	4.408
		lax-creaky	2.988*	.511	.000	1.308	4.667
		harsh	-.069	.169	1.000	-.626	.488
modal	m_whispery	m_breathy	-.381	.123	.073	-.785	.023
		m_lax-creaky	-.581*	.165	.031	-1.125	-.038
		m_harsh	-1.662*	.214	.000	-2.365	-.960
		m_tense	-1.812*	.249	.000	-2.630	-.995
	m_breathy	m_whispery	.381	.123	.073	-.023	.785
		m_lax-creaky	-.200	.091	.437	-.498	.098
		m_harsh	-1.281*	.170	.000	-1.840	-.722
		m_tense	-1.431*	.180	.000	-2.021	-.841
	m_lax-creaky	m_whispery	.581*	.165	.031	.038	1.125
		m_breathy	.200	.091	.437	-.098	.498
		m_harsh	-1.081*	.156	.000	-1.594	-.568
		m_tense	-1.231*	.161	.000	-1.759	-.703
	m_harsh	m_whispery	1.662*	.214	.000	.960	2.365
		m_breathy	1.281*	.170	.000	.722	1.840
		m_lax-creaky	1.081*	.156	.000	.568	1.594
		m_tense	-.150	.130	1.000	-.576	.276
	m_tense	m_whispery	1.812*	.249	.000	.995	2.630
		m_breathy	1.431*	.180	.000	.841	2.021
		m_lax-creaky	1.231*	.161	.000	.703	1.759
		m_harsh	.150	.130	1.000	-.276	.576

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

*. The mean difference is significant at the .05 level.

Bored-Interested

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
loudness	Sphericity Assumed	175.826	4	43.956	70.897	.000	.825
	Greenhouse-Geisser	175.826	1.933	90.944	70.897	.000	.825
Error(loudness)	Sphericity Assumed	37.200	60	.620			
	Greenhouse-Geisser	37.200	29.000	1.283			
voice	Sphericity Assumed	17.227	1	17.227	29.097	.000	.660
	Greenhouse-Geisser	17.227	1.000	17.227	29.097	.000	.660
Error(voice)	Sphericity Assumed	8.880	15	.592			
	Greenhouse-Geisser	8.880	15.000	.592			
loudness*voice	Sphericity Assumed	29.263	4	7.316	16.655	.000	.526
	Greenhouse-Geisser	29.263	1.813	16.137	16.655	.000	.526
Error(loudness *voice)	Sphericity Assumed	26.355	60	.439			
	Greenhouse-Geisser	26.355	27.201	.969			

Bored-Interested: Pairwise Comparisons I

loudness	(I) voice	(J) voice	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
						Lower Bound	Upper Bound
1	whispery	m_whispery	-.544*	.172	.006	-.910	-.177
2	breathy	m_breathy	-.475*	.193	.026	-.886	-.064
3	lax-creaky	m_lax-creaky	-2.256*	.234	.000	-2.755	-1.758
4	harsh	m_harsh	-.319	.373	.406	-1.113	.475
5	tense	m_tense	.313	.183	.108	-.077	.702

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

a. Adjustment for multiple comparisons: Bonferroni.

Bored-Interested: Pairwise Comparisons II

voice	(I) loudness	(J) loudness	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
						Lower Bound	Upper Bound
original	whispery	breathy	-.231	.210	1.000	-.920	.458
		lax-creaky	1.169*	.224	.001	.434	1.903
		harsh	-1.919*	.427	.004	-3.320	-.517
		tense	-2.725*	.271	.000	-3.617	-1.833
	breathy	whispery	.231	.210	1.000	-.458	.920
		lax-creaky	1.400*	.270	.001	.512	2.288
		harsh	-1.688*	.508	.046	-3.356	-.019
		tense	-2.494*	.338	.000	-3.605	-1.383
	lax-creaky	whispery	-1.169*	.224	.001	-1.903	-.434
		breathy	-1.400*	.270	.001	-2.288	-.512
		harsh	-3.088*	.463	.000	-4.608	-1.567
		tense	-3.894*	.303	.000	-4.889	-2.899
	harsh	whispery	1.919*	.427	.004	.517	3.320
		breathy	1.688*	.508	.046	.019	3.356
		lax-creaky	3.088*	.463	.000	1.567	4.608
		tense	-.806*	.231	.033	-1.567	-.046
	tense	whispery	2.725*	.271	.000	1.833	3.617
		breathy	2.494*	.338	.000	1.383	3.605
		lax-creaky	3.894*	.303	.000	2.899	4.889
		harsh	.806*	.231	.033	.046	1.567
modal	m_whispery	m_breathy	-.163	.100	1.000	-.490	.165
		m_lax-creaky	-.544*	.135	.011	-.989	-.099
		m_harsh	-1.694*	.150	.000	-2.187	-1.200
		m_tense	-1.869*	.181	.000	-2.465	-1.273
	m_breathy	m_whispery	.163	.100	1.000	-.165	.490
		m_lax-creaky	-.381*	.084	.004	-.657	-.106
		m_harsh	-1.531*	.121	.000	-1.928	-1.135
		m_tense	-1.706*	.137	.000	-2.156	-1.257
	m_lax-creaky	m_whispery	.544*	.135	.011	.099	.989
		m_breathy	.381*	.084	.004	.106	.657
		m_harsh	-1.150*	.105	.000	-1.496	-.804
		m_tense	-1.325*	.138	.000	-1.778	-.872
	m_harsh	m_whispery	1.694*	.150	.000	1.200	2.187
		m_breathy	1.531*	.121	.000	1.135	1.928
		m_lax-creaky	1.150*	.105	.000	.804	1.496
		m_tense	-.175	.126	1.000	-.588	.238
	m_tense	m_whispery	1.869*	.181	.000	1.273	2.465
		m_breathy	1.706*	.137	.000	1.257	2.156
		m_lax-creaky	1.325*	.138	.000	.872	1.778
		m_harsh	.175	.126	1.000	-.238	.588

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

*. The mean difference is significant at the .05 level.

Intimate-Formal

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
loudness	Sphericity Assumed	299.050	4	74.763	199.086	.000	.930
	Greenhouse-Geisser	299.050	2.542	117.621	199.086	.000	.930
Error(loudness)	Sphericity Assumed	22.532	60	.376			
	Greenhouse-Geisser	22.532	38.137	.591			
voice	Sphericity Assumed	36.005	1	36.005	53.729	.000	.782
	Greenhouse-Geisser	36.005	1.000	36.005	53.729	.000	.782
Error(voice)	Sphericity Assumed	10.052	15	.670			
	Greenhouse-Geisser	10.052	15.000	.670			
loudness*voice	Sphericity Assumed	27.874	4	6.969	19.371	.000	.564
	Greenhouse-Geisser	27.874	2.044	13.637	19.371	.000	.564
Error(loudness *voice)	Sphericity Assumed	21.584	60	.360			
	Greenhouse-Geisser	21.584	30.661	.704			

Intimate-Formal: Pairwise Comparisons I

loudness	(I) voice	(J) voice	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
						Lower Bound	Upper Bound
1	whispery	m_whispery	-1.238*	.129	.000	-1.513	-.962
2	breathy	m_breathy	-1.312*	.099	.000	-1.524	-1.101
3	lax-creaky	m_lax-creaky	-2.175*	.275	.000	-2.761	-1.589
4	harsh	m_harsh	-.013	.311	.968	-.675	.650
5	tense	m_tense	-.006	.255	.981	-.549	.537

Based on estimated marginal means

*. The mean difference is significant at the .05 level

a. Adjustment for multiple comparisons: Bonferroni.

Intimate-Formal: Pairwise Comparisons II

voice	(I) loudness	(J) loudness	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
						Lower Bound	Upper Bound
original	whispery	breathy	-.106	.132	1.000	-.539	.327
		lax-creaky	.294	.319	1.000	-.755	1.343
		harsh	-3.331*	.286	.000	-4.270	-2.392
		tense	-3.669*	.215	.000	-4.375	-2.963
	breathy	whispery	.106	.132	1.000	-.327	.539
		lax-creaky	.400	.297	1.000	-.575	1.375
		harsh	-3.225*	.246	.000	-4.032	-2.418
		tense	-3.562*	.175	.000	-4.137	-2.988
	lax-creaky	whispery	-.294	.319	1.000	-1.343	.755
		breathy	-.400	.297	1.000	-1.375	.575
		harsh	-3.625*	.430	.000	-5.039	-2.211
		tense	-3.963*	.330	.000	-5.047	-2.878
	harsh	whispery	3.331*	.286	.000	2.392	4.270
		breathy	3.225*	.246	.000	2.418	4.032
		lax-creaky	3.625*	.430	.000	2.211	5.039
		tense	-.338	.175	.727	-.912	.237
	tense	whispery	3.669*	.215	.000	2.963	4.375
		breathy	3.562*	.175	.000	2.988	4.137
		lax-creaky	3.963*	.330	.000	2.878	5.047
		harsh	.338	.175	.727	-.237	.912
modal	m_whispery	m_breathy	-.181	.128	1.000	-.603	.240
		m_lax-creaky	-.644*	.099	.000	-.968	-.319
		m_harsh	-2.106*	.156	.000	-2.620	-1.592
		m_tense	-2.438*	.151	.000	-2.933	-1.942
	m_breathy	m_whispery	.181	.128	1.000	-.240	.603
		m_lax-creaky	-.462*	.058	.000	-.654	-.271
		m_harsh	-1.925*	.148	.000	-2.413	-1.437
		m_tense	-2.256*	.167	.000	-2.806	-1.706
	m_lax-creaky	m_whispery	.644*	.099	.000	.319	.968
		m_breathy	.462*	.058	.000	.271	.654
		m_harsh	-1.463*	.131	.000	-1.892	-1.033
		m_tense	-1.794*	.133	.000	-2.231	-1.357
	m_harsh	m_whispery	2.106*	.156	.000	1.592	2.620
		m_breathy	1.925*	.148	.000	1.437	2.413
		m_lax-creaky	1.463*	.131	.000	1.033	1.892
		m_tense	-.331*	.082	.011	-.601	-.062
	m_tense	m_whispery	2.438*	.151	.000	1.942	2.933
		m_breathy	2.256*	.167	.000	1.706	2.806
		m_lax-creaky	1.794*	.133	.000	1.357	2.231
		m_harsh	.331*	.082	.011	.062	.601

Based on estimated marginal means

*. The mean difference is significant at the .05 level

a. Adjustment for multiple comparisons: Bonferroni.

Relaxed-Stressed

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
loudness	Sphericity Assumed	320.936	4	80.234	218.164	.000	.936
	Greenhouse-Geisser	320.936	1.518	211.367	218.164	.000	.936
Error(loudness)	Sphericity Assumed	22.066	60	.368			
	Greenhouse-Geisser	22.066	22.776	.969			
voice	Sphericity Assumed	12.377	1	12.377	21.902	.000	.594
	Greenhouse-Geisser	12.377	1.000	12.377	21.902	.000	.594
Error(voice)	Sphericity Assumed	8.476	15	.565			
	Greenhouse-Geisser	8.476	15.000	.565			
loudness*voice	Sphericity Assumed	73.261	4	18.315	84.393	.000	.849
	Greenhouse-Geisser	73.261	1.910	38.364	84.393	.000	.849
Error(loudness *voice)	Sphericity Assumed	13.021	60	.217			
	Greenhouse-Geisser	13.021	28.644	.455			

Relaxed-Stressed: Pairwise Comparisons I

loudness	(I) voice	(J) voice	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
						Lower Bound	Upper Bound
1	whispery	m_whispery	-.744*	.159	.000	-1.082	-.406
2	breathy	m_breathy	-1.194*	.117	.000	-1.444	-.943
3	lax-creaky	m_lax-creaky	-2.656*	.164	.000	-3.005	-2.308
4	harsh	m_harsh	1.025*	.293	.003	.401	1.649
5	tense	m_tense	.788*	.167	.000	.432	1.143

Based on estimated marginal means

*. The mean difference is significant at the .05 level

a. Adjustment for multiple comparisons: Bonferroni.

Relaxed-Stressed: Pairwise Comparisons II

voice	(I) loudness	(J) loudness	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
						Lower Bound	Upper Bound
original	whispery	breathy	.244	.142	1.000	-.223	.710
		lax-creaky	1.344*	.166	.000	.798	1.889
		harsh	-3.562*	.324	.000	-4.629	-2.496
		tense	-3.569*	.256	.000	-4.411	-2.726
	breathy	whispery	-.244	.142	1.000	-.710	.223
		lax-creaky	1.100*	.131	.000	.668	1.532
		harsh	-3.806*	.272	.000	-4.700	-2.913
		tense	-3.812*	.188	.000	-4.429	-3.196
	lax-creaky	whispery	-1.344*	.166	.000	-1.889	-.798
		breathy	-1.100*	.131	.000	-1.532	-.668
		harsh	-4.906*	.281	.000	-5.829	-3.983
		tense	-4.912*	.183	.000	-5.512	-4.313
	harsh	whispery	3.562*	.324	.000	2.496	4.629
		breathy	3.806*	.272	.000	2.913	4.700
		lax-creaky	4.906*	.281	.000	3.983	5.829
		tense	-.006	.160	1.000	-.531	.518
	tense	whispery	3.569*	.256	.000	2.726	4.411
		breathy	3.812*	.188	.000	3.196	4.429
		lax-creaky	4.912*	.183	.000	4.313	5.512
		harsh	.006	.160	1.000	-.518	.531
modal	m_whispery	m_breathy	-.206	.105	.694	-.553	.140
		m_lax-creaky	-.569*	.122	.003	-.969	-.169
		m_harsh	-1.794*	.223	.000	-2.525	-1.062
		m_tense	-2.037*	.232	.000	-2.800	-1.275
	m_breathy	m_whispery	.206	.105	.694	-.140	.553
		m_lax-creaky	-.362*	.082	.005	-.631	-.094
		m_harsh	-1.588*	.152	.000	-2.087	-1.088
		m_tense	-1.831*	.180	.000	-2.422	-1.241
	m_lax-creaky	m_whispery	.569*	.122	.003	.169	.969
		m_breathy	.362*	.082	.005	.094	.631
		m_harsh	-1.225*	.150	.000	-1.719	-.731
		m_tense	-1.469*	.164	.000	-2.009	-.928
	m_harsh	m_whispery	1.794*	.223	.000	1.062	2.525
		m_breathy	1.588*	.152	.000	1.088	2.087
		m_lax-creaky	1.225*	.150	.000	.731	1.719
		m_tense	-.244	.085	.115	-.522	.034
	m_tense	m_whispery	2.037*	.232	.000	1.275	2.800
		m_breathy	1.831*	.180	.000	1.241	2.422
		m_lax-creaky	1.469*	.164	.000	.928	2.009
		m_harsh	.244	.085	.115	-.034	.522

Based on estimated marginal means

*. The mean difference is significant at the .05 level

a. Adjustment for multiple comparisons: Bonferroni.

Sad-Happy

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
loudness	Sphericity Assumed	162.120	4	40.530	69.319	.000	.822
	Greenhouse-Geisser	162.120	1.693	95.750	69.319	.000	.822
Error(loudness)	Sphericity Assumed	35.082	60	.585			
	Greenhouse-Geisser	35.082	25.397	1.381			
voice	Sphericity Assumed	33.124	1	33.124	72.789	.000	.829
	Greenhouse-Geisser	33.124	1.000	33.124	72.789	.000	.829
Error(voice)	Sphericity Assumed	6.826	15	.455			
	Greenhouse-Geisser	6.826	15.000	.455			
loudness*voice	Sphericity Assumed	36.952	4	9.238	35.173	.000	.701
	Greenhouse-Geisser	36.952	2.176	16.982	35.173	.000	.701
Error(loudness * voice)	Sphericity Assumed	15.758	60	.263			
	Greenhouse-Geisser	15.758	32.639	.483			

Sad-Happy: Pairwise Comparisons I

loudness	(I) voice	(J) voice	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
						Lower Bound	Upper Bound
1	whispery	m_whispery	-1.038*	.106	.000	-1.263	-.812
2	breathy	m_breathy	-1.169*	.105	.000	-1.392	-.945
3	lax-creaky	m_lax-creaky	-2.506*	.224	.000	-2.983	-2.030
4	harsh	m_harsh	-.025	.281	.930	-.625	.575
5	tense	m_tense	.187	.192	.345	-.222	.597

Based on estimated marginal means

*. The mean difference is significant at the .05 level

a. Adjustment for multiple comparisons: Bonferroni.

Sad-Happy: Pairwise Comparisons II

voice	(I) loudness	(J) loudness	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
						Lower Bound	Upper Bound
original	whispery	breathy	-.163	.104	1.000	-.503	.178
		lax-creaky	.856*	.217	.013	.144	1.569
		harsh	-2.506*	.357	.000	-3.678	-1.334
		tense	-2.762*	.295	.000	-3.730	-1.795
	breathy	whispery	.163	.104	1.000	-.178	.503
		lax-creaky	1.019*	.207	.002	.337	1.700
		harsh	-2.344*	.344	.000	-3.476	-1.212
		tense	-2.600*	.278	.000	-3.513	-1.687
	lax-creaky	whispery	-.856*	.217	.013	-1.569	-.144
		breathy	-1.019*	.207	.002	-1.700	-.337
		harsh	-3.362*	.401	.000	-4.680	-2.045
		tense	-3.619*	.380	.000	-4.869	-2.369
	harsh	whispery	2.506*	.357	.000	1.334	3.678
		breathy	2.344*	.344	.000	1.212	3.476
		lax-creaky	3.362*	.401	.000	2.045	4.680
		tense	-.256	.204	1.000	-.927	.414
	tense	whispery	2.762*	.295	.000	1.795	3.730
		breathy	2.600*	.278	.000	1.687	3.513
		lax-creaky	3.619*	.380	.000	2.369	4.869
		harsh	.256	.204	1.000	-.414	.927
modal	m_whispery	m_breathy	-.294*	.077	.017	-.547	-.040
		m_lax-creaky	-.612*	.106	.000	-.962	-.263
		m_harsh	-1.494*	.181	.000	-2.087	-.900
		m_tense	-1.537*	.178	.000	-2.122	-.953
	m_breathy	m_whispery	.294*	.077	.017	.040	.547
		m_lax-creaky	-.319	.106	.089	-.668	.030
		m_harsh	-1.200*	.156	.000	-1.712	-.688
		m_tense	-1.244*	.146	.000	-1.725	-.763
	m_lax-creaky	m_whispery	.612*	.106	.000	.263	.962
		m_breathy	.319	.106	.089	-.030	.668
		m_harsh	-.881*	.160	.001	-1.408	-.355
		m_tense	-.925*	.175	.001	-1.502	-.348
	m_harsh	m_whispery	1.494*	.181	.000	.900	2.087
		m_breathy	1.200*	.156	.000	.688	1.712
		m_lax-creaky	.881*	.160	.001	.355	1.408
		m_tense	-.044	.081	1.000	-.309	.221
	m_tense	m_whispery	1.537*	.178	.000	.953	2.122
		m_breathy	1.244*	.146	.000	.763	1.725
		m_lax-creaky	.925*	.175	.001	.348	1.502
		m_harsh	.044	.081	1.000	-.221	.309

Based on estimated marginal means

*. The mean difference is significant at the .05 level

a. Adjustment for multiple comparisons: Bonferroni.

Scared-Fearless

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
loudness	Sphericity Assumed	47.068	4	11.767	4.245	.004	.221
	Greenhouse-Geisser	47.068	1.277	36.858	4.245	.045	.221
Error(loudness)	Sphericity Assumed	166.316	60	2.772			
	Greenhouse-Geisser	166.316	19.155	8.683			
voice	Sphericity Assumed	2.093	1	2.093	2.246	.155	.130
	Greenhouse-Geisser	2.093	1.000	2.093	2.246	.155	.130
Error(voice)	Sphericity Assumed	13.976	15	.932			
	Greenhouse-Geisser	13.976	15.000	.932			
loudness*voice	Sphericity Assumed	4.561	4	1.140	1.398	.246	.085
	Greenhouse-Geisser	4.561	1.774	2.571	1.398	.263	.085
Error(loudness *voice)	Sphericity Assumed	48.955	60	.816			
	Greenhouse-Geisser	48.955	26.605	1.840			

Scared-Fearless: Pairwise Comparisons I

loudness	(I) voice	(J) voice	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
						Lower Bound	Upper Bound
1	whispery	m_whispery	-.656*	.167	.001	-1.012	-.300
2	breathy	m_breathy	-.262	.200	.210	-.690	.165
3	lax-creaky	m_lax-creaky	.362	.475	.457	-.650	1.375
4	harsh	m_harsh	-.413	.395	.313	-1.255	.430
5	tense	m_tense	-.175	.272	.530	-.755	.405

Based on estimated marginal means

*. The mean difference is significant at the .05 level

a. Adjustment for multiple comparisons: Bonferroni.

Scared-Fearless: Pairwise Comparisons II

voice	(I) loudness	(J) loudness	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
						Lower Bound	Upper Bound
original	whispery	breathy	-.600*	.139	.006	-1.056	-.144
		lax-creaky	-1.381	.434	.062	-2.808	.045
		harsh	-1.388	.675	.577	-3.606	.831
		tense	-1.713	.643	.178	-3.827	.402
	breathy	whispery	.600*	.139	.006	.144	1.056
		lax-creaky	-.781	.366	.498	-1.985	.422
		harsh	-.788	.679	1.000	-3.019	1.444
		tense	-1.112	.662	1.000	-3.288	1.063
	lax-creaky	whispery	1.381	.434	.062	-.045	2.808
		breathy	.781	.366	.498	-.422	1.985
		harsh	-.006	.865	1.000	-2.848	2.836
		tense	-.331	.843	1.000	-3.100	2.438
	harsh	whispery	1.388	.675	.577	-.831	3.606
		breathy	.788	.679	1.000	-1.444	3.019
		lax-creaky	.006	.865	1.000	-2.836	2.848
		tense	-.325	.251	1.000	-1.148	.498
	tense	whispery	1.713	.643	.178	-.402	3.827
		breathy	1.112	.662	1.000	-1.063	3.288
		lax-creaky	.331	.843	1.000	-2.438	3.100
		harsh	.325	.251	1.000	-.498	1.148
modal	m_whispery	m_breathy	-.206	.092	.413	-.510	.097
		m_lax-creaky	-.362	.159	.378	-.885	.160
		m_harsh	-1.144	.358	.060	-2.320	.032
		m_tense	-1.231	.459	.170	-2.739	.277
	m_breathy	m_whispery	.206	.092	.413	-.097	.510
		m_lax-creaky	-.156	.102	1.000	-.490	.178
		m_harsh	-.938	.300	.069	-1.922	.047
		m_tense	-1.025	.402	.222	-2.345	.295
	m_lax-creaky	m_whispery	.362	.159	.378	-.160	.885
		m_breathy	.156	.102	1.000	-.178	.490
		m_harsh	-.781	.266	.101	-1.654	.091
		m_tense	-.869	.359	.285	-2.047	.309
	m_harsh	m_whispery	1.144	.358	.060	-.032	2.320
		m_breathy	.938	.300	.069	-.047	1.922
		m_lax-creaky	.781	.266	.101	-.091	1.654
		m_tense	-.087	.141	1.000	-.551	.376
	m_tense	m_whispery	1.231	.459	.170	-.277	2.739
		m_breathy	1.025	.402	.222	-.295	2.345
		m_lax-creaky	.869	.359	.285	-.309	2.047
		m_harsh	.087	.141	1.000	-.376	.551

Based on estimated marginal means

*. The mean difference is significant at the .05 level

a. Adjustment for multiple comparisons: Bonferroni.

10 Appendix 10

Shown is the raters' percent agreement in each of the subtests in Loudness experiment 1 (Chapter 5).

The data shows percentage of the raters that assigned ratings above +/- 1 to a particular voice quality stimulus. Ratings within the range of -1 to +1 are interpreted here as suggestive of 'no affect' signalled. Cases in which 75% of the raters or more agreed in voice to affect association are shown in bold type.

Table 10.1. *Raters' percent agreement in Loudness experiment 1. Cases of 75% agreement and above in voice to affect association are shown in bold type.*

Apologetic-Indignant			
	apologetic	indignant	no affect
whispery	**94**	6	0
breathy	**100**	0	0
lax-creaky	63	38	0
harsh	0	**88**	13
tense	0	**100**	0
modal	19	50	31
m_whispery	69	6	19
m_breathy	38	6	56
m_lax-creaky	44	13	44
m_harsh	13	**88**	0
m_tense	13	**88**	0
Bored-Interested			
	bored	interested	no affect
whispery	**88**	6	6
breathy	**75**	13	13
lax-creaky	**94**	6	0
harsh	19	56	25
tense	0	**100**	0
modal	13	63	25
m_whispery	**75**	6	19
m_breathy	**81**	6	13
m_lax-creaky	44	19	38
m_harsh	6	**88**	6
m_tense	0	**100**	0
Intimate-Formal			
	intimate	formal	no affect
whispery	**100**	0	0
breathy	**100**	0	0
lax-creaky	**94**	0	6
harsh	6	**75**	19
tense	0	**94**	6
modal	0	**94**	6
m_whispery	50	6	44
m_breathy	38	6	56
m_lax-creaky	25	19	56
m_harsh	0	**94**	6
m_tense	0	**100**	0
Relaxed-Stressed			
	relaxed	stressed	no affect
whispery	**94**	0	6
breathy	**94**	0	6
lax-creaky	**100**	0	0
harsh	0	**88**	13
tense	0	**94**	6
modal	0	63	38
m_whispery	63	13	25
m_breathy	56	6	38
m_lax-creaky	31	25	38
m_harsh	0	**81**	19
m_tense	0	**88**	13

Table 10.2. *Raters' percent agreement in Loudness experiment 1 (cont.). Cases of 75% agreement and above in voice to affect association are shown in bold type.*

Sad-Happy			
	sad	happy	no affect
whispery	**100**	0	0
breathy	**100**	0	0
lax-creaky	**100**	0	0
harsh	25	50	25
tense	13	**75**	13
modal	6	50	44
m_whispery	56	6	38
m_breathy	38	0	63
m_lax-creaky	25	31	44
m_harsh	6	**75**	19
m_tense	0	**75**	25
Scared-Fearless			
	scared	fearless	no affect
whispery	**75**	6	19
breathy	63	19	19
lax-creaky	56	44	0
harsh	38	38	25
tense	38	56	6
modal	19	44	38
m_whispery	44	6	50
m_breathy	38	19	44
m_lax-creaky	31	6	63
m_harsh	31	63	6
m_tense	25	75	0

11 Appendix 11

Raters' agreement in Loudness experiment 1 (Chapter 5).

Pearson's *r* correlation coefficients were calculated showing correlation of each individual rater with the group mean for a particular subtest. This allows one to ascertain what raters within each group of listeners were found to assign ratings very different to those of a group as a whole.

These coefficients are presented in Table 10.1 below. Cases of Pearson's *r* below 0.5 (suggestive of substantial difference between individual raters and a group as a whole) are shown in bold type.

Table 11.1. *Pearson's* r *showing correlation between the ratings of individual raters and the group mean.*

	Pearson's r correlation with the group mean per test					
Subject	**Apologetic-Indignant**	**Bored-Interested**	**Intimate-Formal**	**Relaxed-Stressed**	**Sad-Happy**	**Scared-Fearless**
S_01	0.88	0.89	0.97	0.91	0.97	**0.05**
S_02	0.86	0.88	0.96	0.99	0.96	0.83
S_03	0.85	0.95	0.92	0.94	0.95	0.88
S_04	0.93	0.97	0.92	0.98	0.95	0.50
S_05	0.98	0.90	0.91	0.99	0.99	0.83
S_06	**0.32**	0.98	0.98	0.99	0.96	0.66
S_07	0.99	0.92	0.95	0.93	0.92	0.79
S_08	0.85	0.69	0.91	0.97	0.70	**-0.26**
S_09	0.98	0.94	0.95	0.99	0.93	**-0.09**
S_10	0.79	0.94	0.96	0.97	0.95	**-0.70**
S_11	0.73	0.96	0.90	0.89	0.92	0.63
S_12	0.92	0.93	0.88	0.97	0.96	**-0.36**
S_13	0.95	**0.49**	0.88	0.96	0.85	0.77
S_14	0.93	0.68	0.97	0.99	0.99	0.93
S_15	0.97	0.89	0.98	0.92	0.96	0.84
S_16	0.95	0.91	0.96	0.94	0.53	0.70

12 Appendix 12

Raters' agreement in Loudness experiment 1 (Chapter 5).

Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for each individual rater, for each stimulus, for each subtest. Each dot in the plots corresponds to the data for one rater. Group mean and standard deviation are shown as a pink dot.

The data allow one to ascertain the extent of agreement/disagreement the raters demonstrated in rating a particular stimulus.

The data also allow one to infer the degree of each rater's confidence in voice to affect association. Given the range of the rating scale, standard deviation of 1 and below (across 10 randomisations) suggests relatively confident voice to affect association.

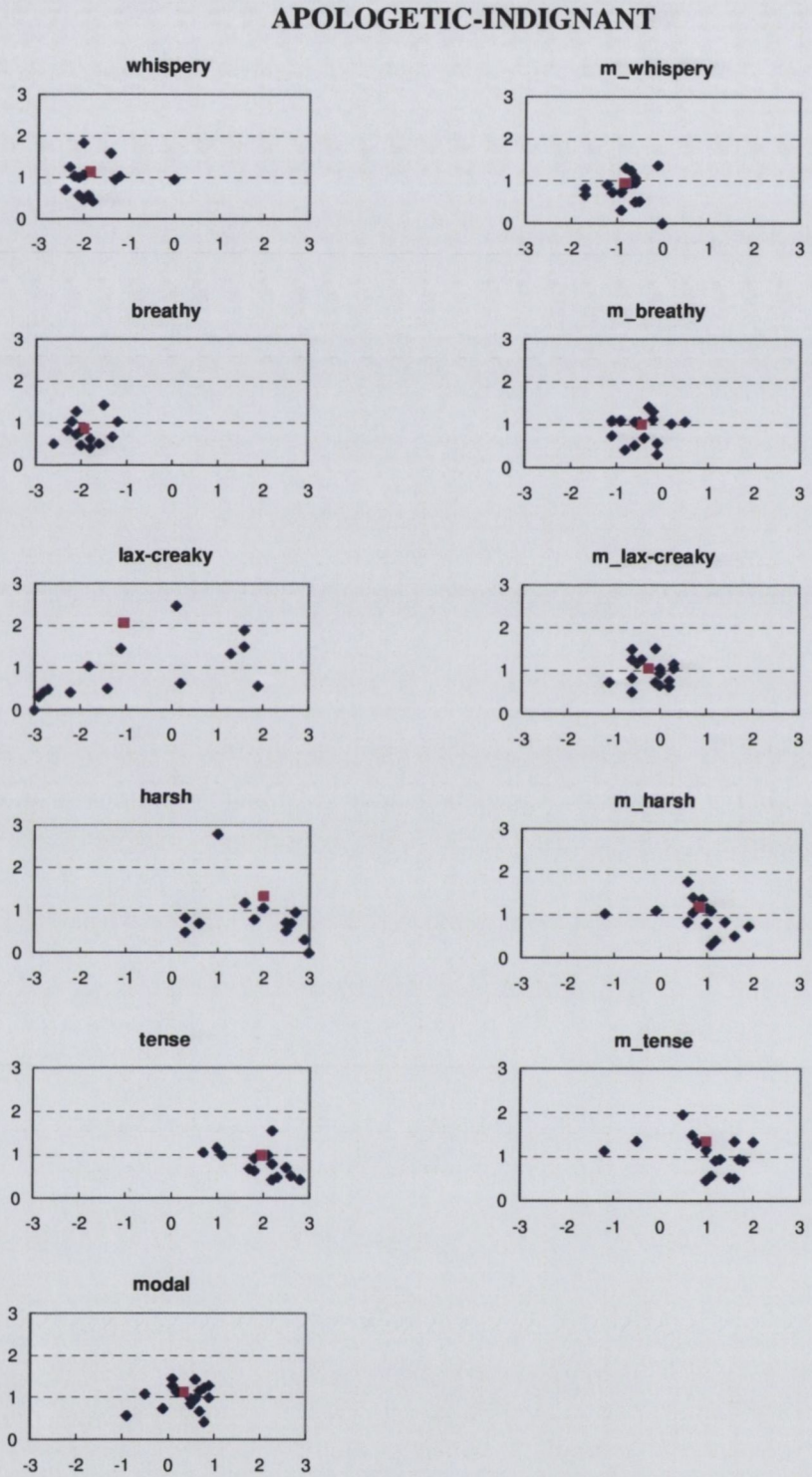


Figure 12.1. *Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for individual raters. Each dot in the plots corresponds to the data for one rater. Group mean and standard deviation are shown as a pink dot.*

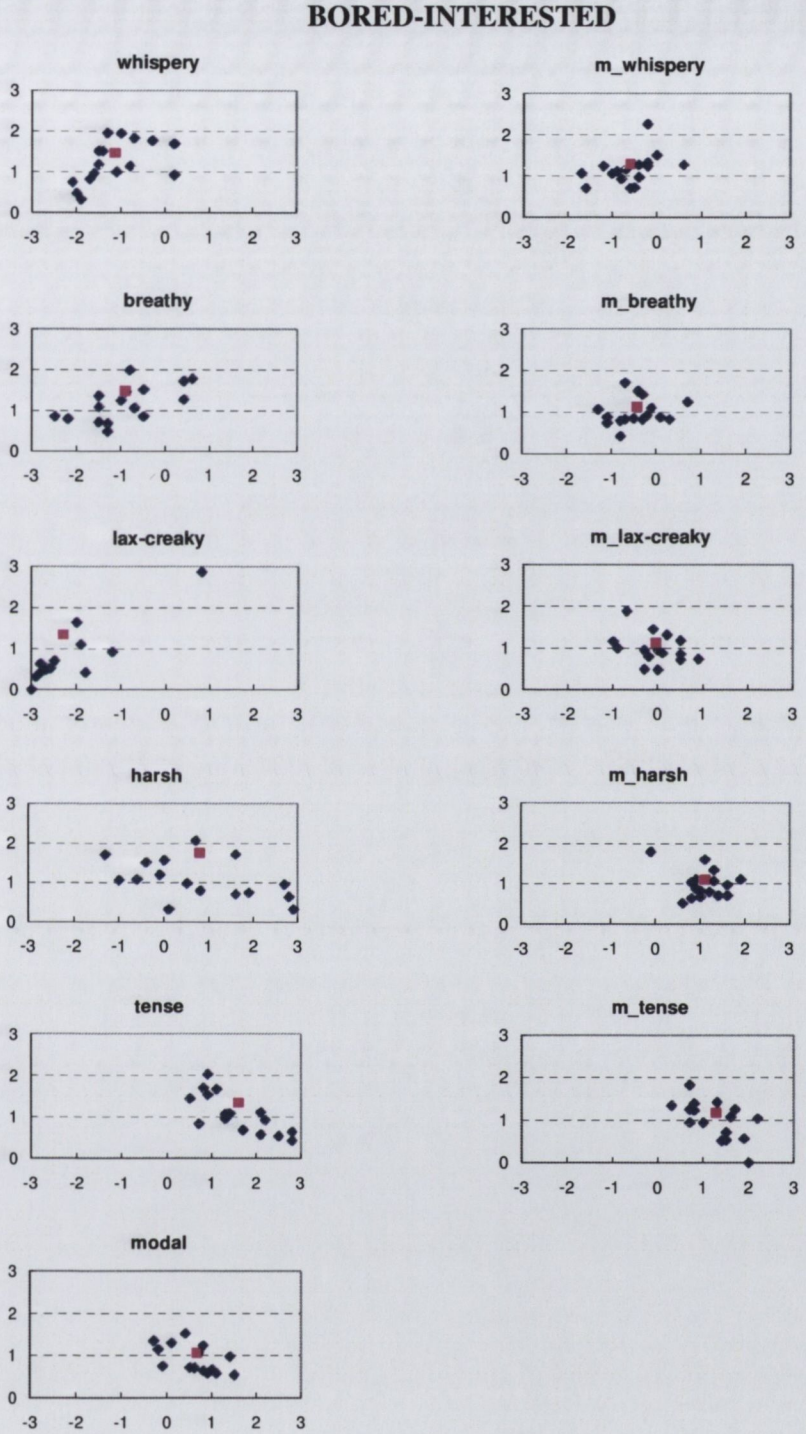


Figure 12.2. *Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for individual raters. Each dot in the plots corresponds to the data for one rater. Group mean and standard deviation are shown as a pink dot.*

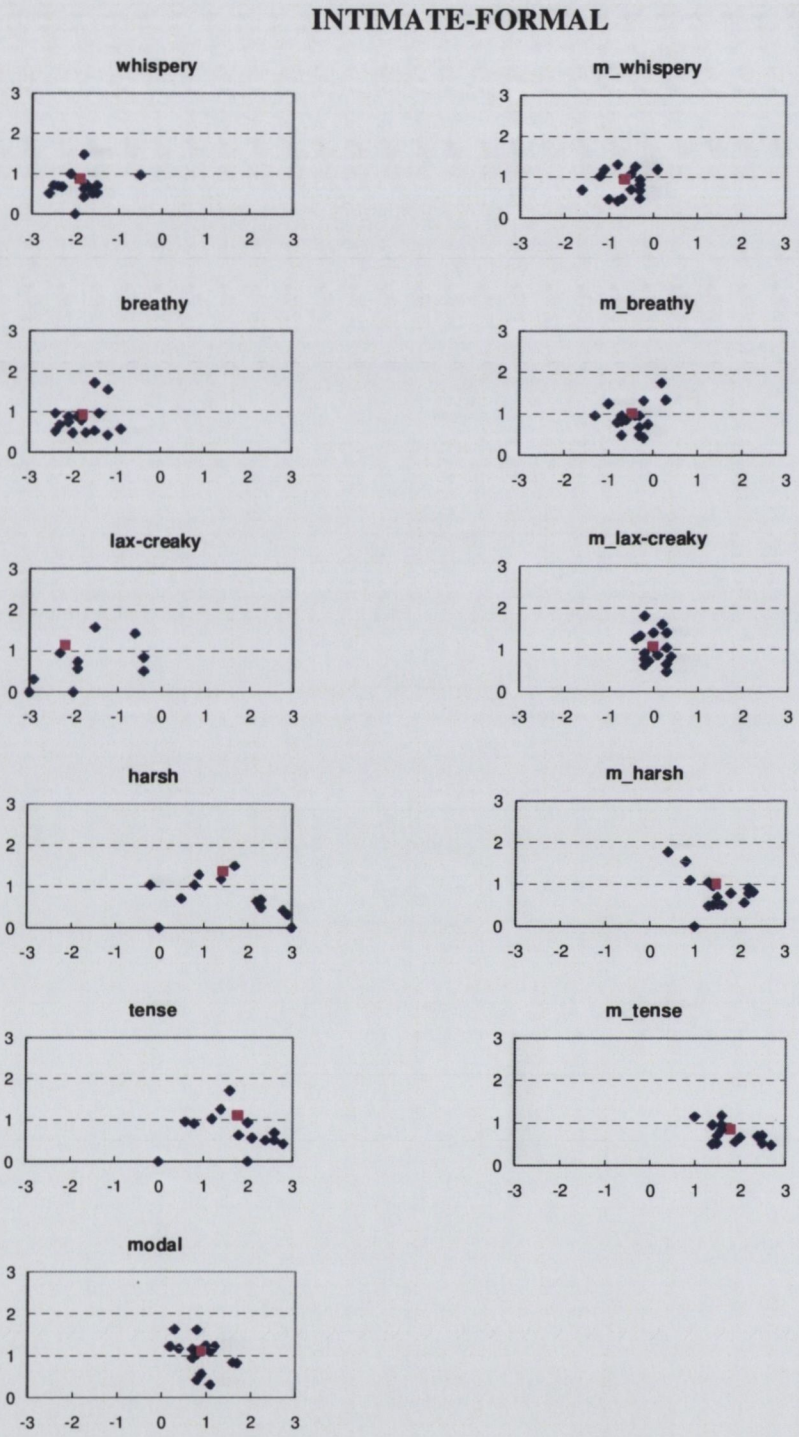


Table 12.3. *Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for individual raters. Each dot in the plots corresponds to the data for one rater. Group mean and standard deviation are shown as a pink dot.*

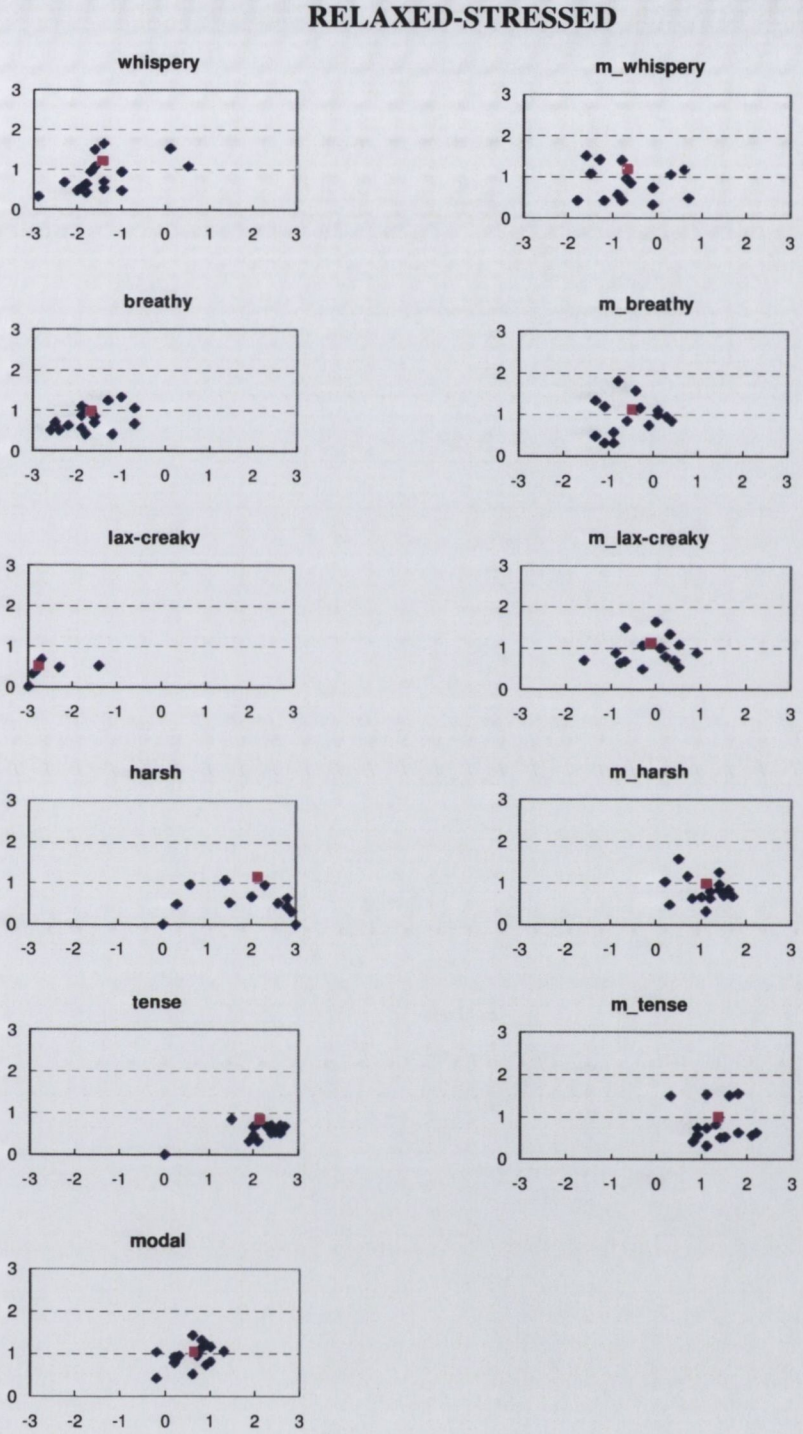


Figure 12.4. *Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for individual raters. Each dot in the plots corresponds to the data for one rater. Group mean and standard deviation are shown as a pink dot.*

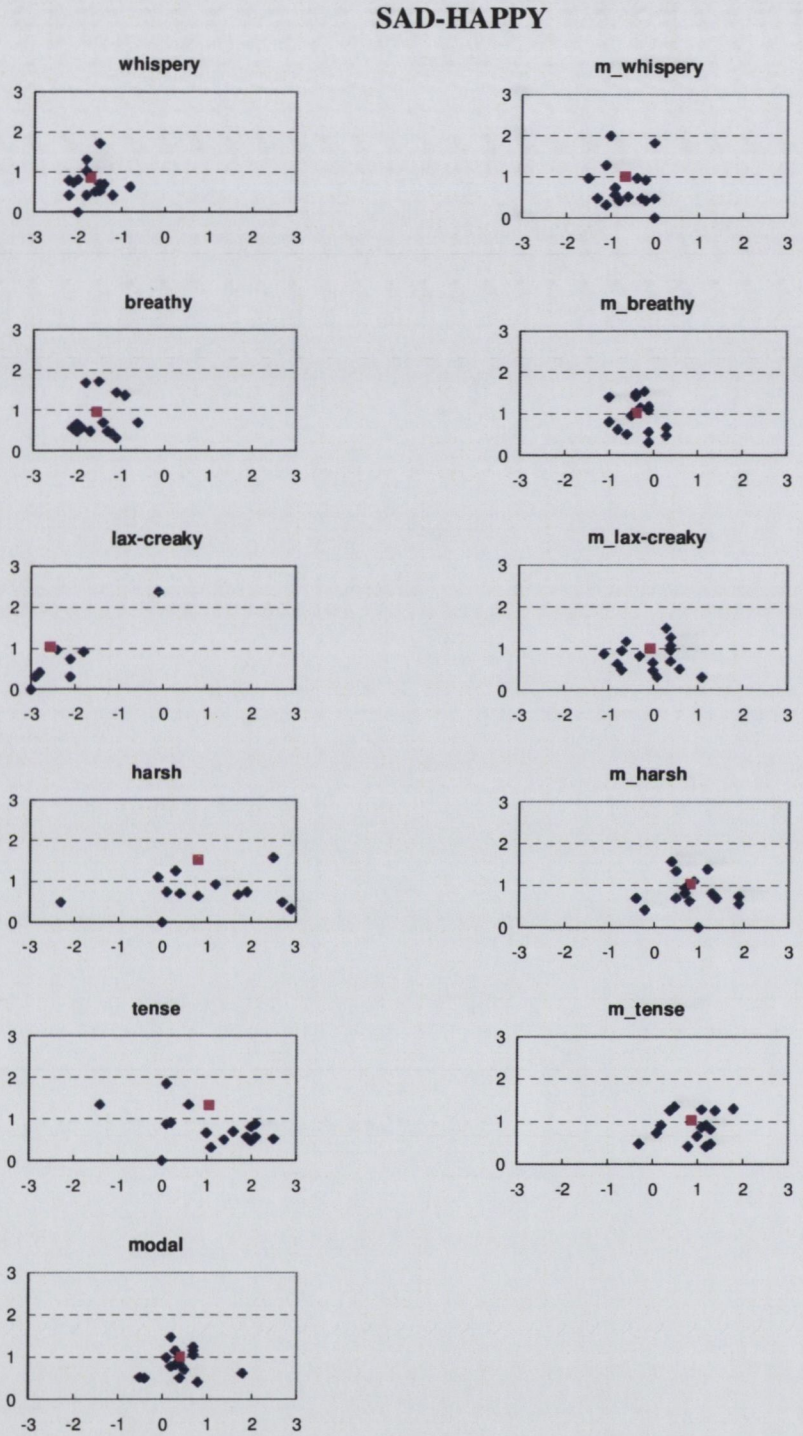


Figure 12.5. *Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for individual raters. Each dot in the plots corresponds to the data for one rater. Group mean and standard deviation are shown as a pink dot.*

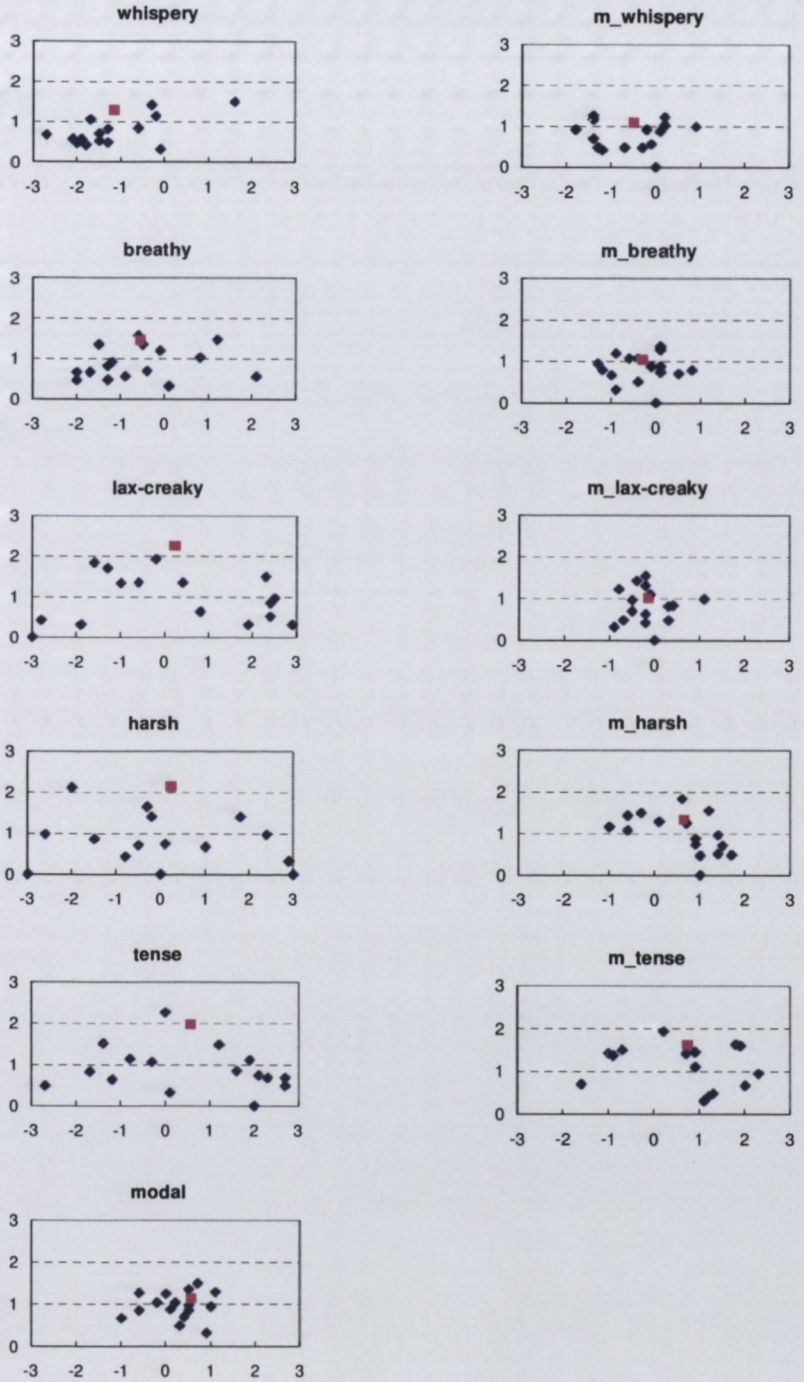


Figure 12.6. *Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for individual raters. Each dot in the plots corresponds to the data for one rater. Group mean and standard deviation are shown as a pink dot.*

13 Appendix 13

Ratings obtained in Loudness experiment 2 (Chapter 5).

In the box plots, mean (dot), median (line), interquartile range (box) and extreme values (whiskers) are shown for the loudness normalised stimuli (Series M) and their quieter (Series Q) and louder (Series L) versions.

Red rectangles highlight the cases where affective ratings yielded by the stimuli were above 1, corresponding to mild to strong affective colouring.

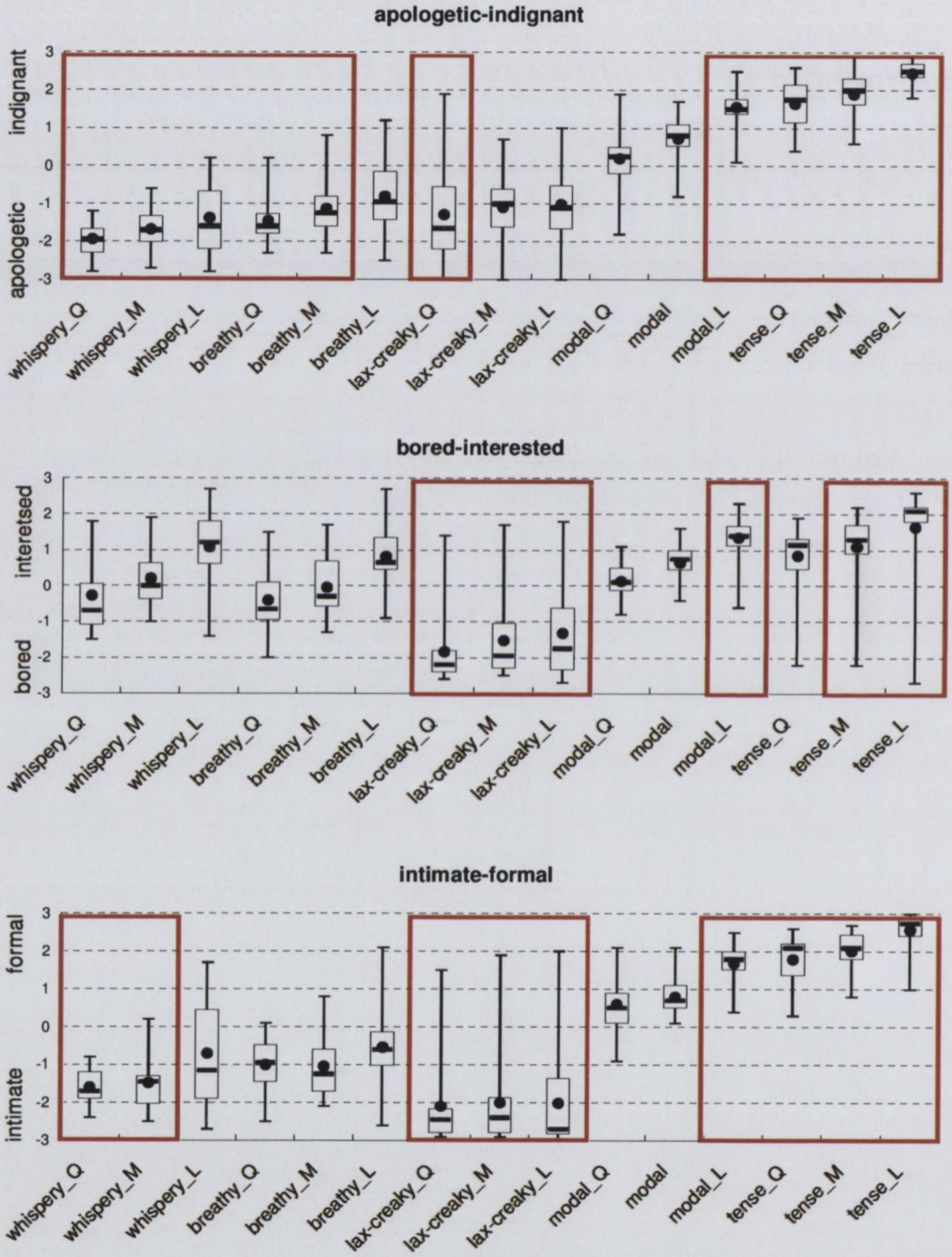


Figure 13.1. *Responses obtained for the loudness normalised stimuli and their quieter and louder versions for each subtest. The plots show interquartile range (box), mean (filled circle), median (horizontal line in box) and extreme values (whiskers).*

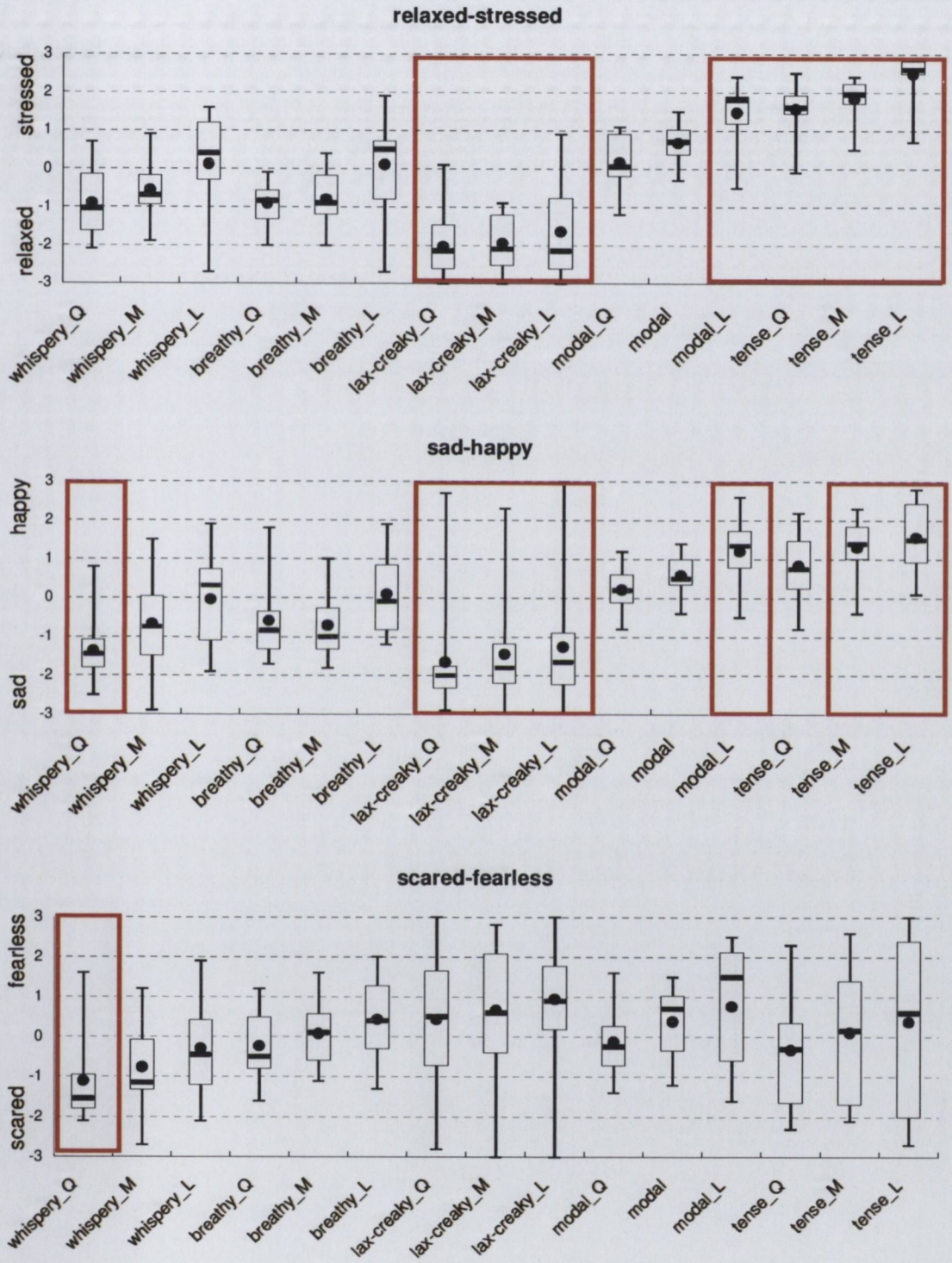


Figure 13.2. *Responses obtained for the loudness normalised stimuli and their quieter and louder versions for each subtest. The plots show interquartile range (box), mean (filled circle), median (horizontal line in box) and extreme values (whiskers).*

14 Appendix 14

Loudness experiment 2 ANOVA

Apologetic-Indignant

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
voice	Sphericity Assumed	468.500	4	117.125	75.495	.000	.834
	Greenhouse-Geisser	468.500	2.357	198.791	75.495	.000	.834
Error(voice)	Sphericity Assumed	93.085	60	1.551			
	Greenhouse-Geisser	93.085	35.351	2.633			
loudness	Sphericity Assumed	21.960	2	10.980	22.215	.000	.597
	Greenhouse-Geisser	21.960	1.270	17.294	22.215	.000	.597
Error (loudness)	Sphericity Assumed	14.828	30	.494			
	Greenhouse-Geisser	14.828	19.047	.778			
voice * loudness	Sphericity Assumed	5.641	8	.705	3.850	.000	.204
	Greenhouse-Geisser	5.641	4.289	1.315	3.850	.006	.204
Error (voice* loudness)	Sphericity Assumed	21.978	120	.183			
	Greenhouse-Geisser	21.978	64.341	.342			

Apologetic-Indignant: Pairwise Comparisons

voice	(I) loudness	(J) loudness	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
						Lower Bound	Upper Bound
whispery	Q	M	-.269	.136	.202	-.636	.098
		L	-.569*	.211	.050	-1.137	.000
	M	Q	.269	.136	.202	-.098	.636
		L	-.300	.165	.269	-.745	.145
breathy	Q	M	-.331	.215	.433	-.910	.248
		L	-.656	.255	.064	-1.344	.031
	M	Q	.331	.215	.433	-.248	.910
		L	-.325	.137	.094	-.693	.043
lax-creaky	Q	M	-.206	.158	.632	-.631	.219
		L	-.281	.207	.583	-.839	.276
	M	Q	.206	.158	.632	-.219	.631
		L	-.075	.121	1.000	-.400	.250
modal	Q	M	-.537	.203	.056	-1.086	.011
		L	-1.375*	.210	.000	-1.941	-.809
	M	Q	.537	.203	.056	-.011	1.086
		L	-.838*	.149	.000	-1.238	-.437
tense	Q	M	-.256	.133	.221	-.615	.103
		L	-.812*	.151	.000	-1.219	-.406
	M	Q	.256	.133	.221	-.103	.615
		L	-.556*	.096	.000	-.814	-.298

Based on estimated marginal means

*. The mean difference is significant at the .05 level

a. Adjustment for multiple comparisons: Bonferroni.

Bored-Interested

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
voice	Sphericity Assumed	209.549	4	52.387	19.653	.000	.567
	Greenhouse-Geisser	209.549	1.643	127.563	19.653	.000	.567
Error (voice)	Sphericity Assumed	159.933	60	2.666			
	Greenhouse-Geisser	159.933	24.641	6.491			
loudness	Sphericity Assumed	42.500	2	21.250	44.222	.000	.747
	Greenhouse-Geisser	42.500	1.220	34.843	44.222	.000	.747
Error (loudness)	Sphericity Assumed	14.416	30	.481			
	Greenhouse-Geisser	14.416	18.296	.788			
voice * loudness	Sphericity Assumed	4.562	8	.570	4.871	.000	.245
	Greenhouse-Geisser	4.562	5.501	.829	4.871	.000	.245
Error (voice* loudness)	Sphericity Assumed	14.049	120	.117			
	Greenhouse-Geisser	14.049	82.510	.170			

Bored-Interested: Pairwise Comparisons

voice	(I) loudness	(J) loudness	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
						Lower Bound	Upper Bound
whispery	Q	M	-.475*	.109	.002	-.769	-.181
		L	-1.350*	.205	.000	-1.903	-.797
	M	Q	.475*	.109	.002	.181	.769
		L	-.875*	.155	.000	-1.291	-.459
breathy	Q	M	-.356*	.106	.013	-.643	-.070
		L	-1.225*	.203	.000	-1.771	-.679
	M	Q	.356*	.106	.013	.070	.643
		L	-.869*	.155	.000	-1.287	-.450
lax-creaky	Q	M	-.313	.137	.112	-.681	.056
		L	-.513*	.177	.034	-.990	-.035
	M	Q	.313	.137	.112	-.056	.681
		L	-.200	.132	.449	-.555	.155
modal	Q	M	-.512*	.166	.022	-.959	-.066
		L	-1.213*	.183	.000	-1.707	-.718
	M	Q	.512*	.166	.022	.066	.959
		L	-.700*	.153	.001	-1.111	-.289
tense	Q	M	-.250	.108	.106	-.541	.041
		L	-.800*	.137	.000	-1.169	-.431
	M	Q	.250	.108	.106	-.041	.541
		L	-.550*	.138	.004	-.921	-.179

Based on estimated marginal means

*. The mean difference is significant at the .05 level

a. Adjustment for multiple comparisons: Bonferroni.

Intimate-Formal

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
voice	Sphericity Assumed	570.216	4	142.554	76.233	.000	.836
	Greenhouse-Geisser	570.216	2.480	229.888	76.233	.000	.836
Error(voice)	Sphericity Assumed	112.198	60	1.870			
	Greenhouse-Geisser	112.198	37.206	3.016			
loudness	Sphericity Assumed	19.965	2	9.982	15.984	.000	.516
	Greenhouse-Geisser	19.965	1.156	17.277	15.984	.001	.516
Error (loudness)	Sphericity Assumed	18.735	30	.625			
	Greenhouse-Geisser	18.735	17.333	1.081			
voice * loudness	Sphericity Assumed	6.184	8	.773	3.863	.000	.205
	Greenhouse-Geisser	6.184	4.187	1.477	3.863	.006	.205
Error (voice* loudness)	Sphericity Assumed	24.010	120	.200			
	Greenhouse-Geisser	24.010	62.799	.382			

Intimate-Formal: Pairwise Comparisons

voice	(I) loudness	(J) loudness	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
						Lower Bound	Upper Bound
whispery	Q	M	-.131	.128	.968	-.477	.215
		L	-.894*	.302	.029	-1.708	-.080
	M	Q	.131	.128	.968	-.215	.477
		L	-.762	.286	.052	-1.532	.007
breathy	Q	M	.044	.185	1.000	-.455	.543
		L	-.469	.268	.301	-1.190	.252
	M	Q	-.044	.185	1.000	-.543	.455
		L	-.512*	.179	.036	-.996	-.029
lax-creaky	Q	M	-.094	.138	1.000	-.465	.277
		L	-.081	.161	1.000	-.516	.353
	M	Q	.094	.138	1.000	-.277	.465
		L	.013	.132	1.000	-.343	.368
modal	Q	M	-.200	.155	.653	-.619	.219
		L	-1.100*	.219	.000	-1.689	-.511
	M	Q	.200	.155	.653	-.219	.619
		L	-.900*	.139	.000	-1.274	-.526
tense	Q	M	-.212	.114	.243	-.519	.094
		L	-.769*	.162	.001	-1.206	-.331
	M	Q	.212	.114	.243	-.094	.519
		L	-.556*	.118	.001	-.875	-.238

Based on estimated marginal means

*. The mean difference is significant at the .05 level

a. Adjustment for multiple comparisons: Bonferroni.

Relaxed-Stressed

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
voice	Sphericity Assumed	408.602	4	102.150	85.883	.000	.851
	Greenhouse-Geisser	408.602	2.350	173.903	85.883	.000	.851
Error(voice)	Sphericity Assumed	71.365	60	1.189			
	Greenhouse-Geisser	71.365	35.244	2.025			
loudness	Sphericity Assumed	37.177	2	18.589	17.093	.000	.533
	Greenhouse-Geisser	37.177	1.083	34.330	17.093	.001	.533
Error (loudness)	Sphericity Assumed	32.625	30	1.088			
	Greenhouse-Geisser	32.625	16.244	2.008			
voice * loudness	Sphericity Assumed	4.051	8	.506	3.244	.002	.178
	Greenhouse-Geisser	4.051	4.691	.864	3.244	.012	.178
Error (voice* loudness)	Sphericity Assumed	18.727	120	.156			
	Greenhouse-Geisser	18.727	70.359	.266			

Relaxed-Stressed: Pairwise Comparisons

voice	(I) loudness	(J) loudness	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
						Lower Bound	Upper Bound
whispery	Q	M	-.337	.146	.105	-.730	.055
		L	-1.019*	.324	.020	-1.892	-.146
	M	Q	.337	.146	.105	-.055	.730
		L	-.681*	.245	.042	-1.343	-.020
breathy	Q	M	-.106	.165	1.000	-.551	.338
		L	-1.012*	.307	.015	-1.839	-.186
	M	Q	.106	.165	1.000	-.338	.551
		L	-.906*	.204	.001	-1.457	-.356
lax-creaky	Q	M	.000	.156	1.000	-.419	.419
		L	-.387	.262	.478	-1.092	.317
	M	Q	.000	.156	1.000	-.419	.419
		L	-.387	.192	.184	-.904	.129
modal	Q	M	-.506*	.155	.015	-.922	-.090
		L	-1.306*	.257	.000	-1.999	-.613
	M	Q	.506*	.155	.015	.090	.922
		L	-.800*	.152	.000	-1.209	-.391
tense	Q	M	-.275	.131	.159	-.628	.078
		L	-.925*	.161	.000	-1.359	-.491
	M	Q	.275	.131	.159	-.078	.628
		L	-.650*	.081	.000	-.867	-.433

Based on estimated marginal means

*. The mean difference is significant at the .05 level

a. Adjustment for multiple comparisons: Bonferroni.

Sad-Happy

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
voice	Sphericity Assumed	225.647	4	56.412	22.196	.000	.597
	Greenhouse-Geisser	225.647	2.095	107.726	22.196	.000	.597
Error(voice)	Sphericity Assumed	152.494	60	2.542			
	Greenhouse-Geisser	152.494	31.420	4.853			
loudness	Sphericity Assumed	28.398	2	14.199	27.670	.000	.648
	Greenhouse-Geisser	28.398	1.192	23.827	27.670	.000	.648
Error (loudness)	Sphericity Assumed	15.394	30	.513			
	Greenhouse-Geisser	15.394	17.877	.861			
voice * loudness	Sphericity Assumed	5.677	8	.710	3.348	.002	.182
	Greenhouse-Geisser	5.677	4.397	1.291	3.348	.012	.182
Error (voice* loudness)	Sphericity Assumed	25.437	120	.212			
	Greenhouse-Geisser	25.437	65.954	.386			

Sad-Happy: Pairwise Comparisons

voice	(I) loudness	(J) loudness	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
						Lower Bound	Upper Bound
whispery	Q	M	-.694*	.174	.003	-1.161	-.226
		L	-1.312*	.253	.000	-1.994	-.631
	M	Q	.694*	.174	.003	.226	1.161
		L	-.619*	.181	.012	-1.108	-.130
breathy	Q	M	.119	.160	1.000	-.313	.550
		L	-.694*	.240	.033	-1.339	-.048
	M	Q	-.119	.160	1.000	-.550	.313
		L	-.812*	.200	.003	-1.351	-.274
lax-creaky	Q	M	-.275	.128	.144	-.619	.069
		L	-.463	.250	.252	-1.136	.211
	M	Q	.275	.128	.144	-.069	.619
		L	-.188	.160	.780	-.619	.244
modal	Q	M	-.375*	.111	.013	-.675	-.075
		L	-.994*	.194	.000	-1.516	-.471
	M	Q	.375*	.111	.013	.075	.675
		L	-.619*	.152	.003	-1.027	-.210
tense	Q	M	-.469*	.143	.015	-.853	-.084
		L	-.725*	.187	.005	-1.229	-.221
	M	Q	.469*	.143	.015	.084	.853
		L	-.256	.165	.426	-.702	.189

Based on estimated marginal means

*. The mean difference is significant at the .05 level

a. Adjustment for multiple comparisons: Bonferroni.

Scared-Fearless

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
voice	Sphericity Assumed	49.911	4	12.478	2.237	.076	.130
	Greenhouse-Geisser	49.911	1.361	36.677	2.237	.145	.130
Error(voice)	Sphericity Assumed	334.653	60	5.578			
	Greenhouse-Geisser	334.653	20.412	16.395			
loudness	Sphericity Assumed	19.811	2	9.905	12.025	.000	.445
	Greenhouse-Geisser	19.811	1.069	18.538	12.025	.003	.445
Error (loudness)	Sphericity Assumed	24.712	30	.824			
	Greenhouse-Geisser	24.712	16.030	1.542			
voice * loudness	Sphericity Assumed	.857	8	.107	.511	.846	.033
	Greenhouse-Geisser	.857	3.623	.236	.511	.710	.033
Error (voice* loudness)	Sphericity Assumed	25.147	120	.210			
	Greenhouse-Geisser	25.147	54.347	.463			

Scared-Fearless: Pairwise Comparisons

voice	(I) loudness	(J) loudness	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
						Lower Bound	Upper Bound
whispery	Q	M	-.287	.144	.192	-.675	.100
		L	-.762*	.282	.049	-1.523	-.002
	M	Q	.287	.144	.192	-.100	.675
		L	-.475	.263	.274	-1.184	.234
breathy	Q	M	-.306*	.099	.023	-.574	-.039
		L	-.669	.254	.056	-1.352	.014
	M	Q	.306*	.099	.023	.039	.574
		L	-.362	.229	.403	-.980	.255
lax-creaky	Q	M	-.231	.114	.181	-.538	.076
		L	-.506	.194	.058	-1.028	.015
	M	Q	.231	.114	.181	-.076	.538
		L	-.275	.122	.119	-.604	.054
modal	Q	M	-.506*	.131	.004	-.858	-.155
		L	-.888*	.302	.030	-1.700	-.075
	M	Q	.506*	.131	.004	.155	.858
		L	-.381	.238	.392	-1.024	.261
tense	Q	M	-.438*	.111	.004	-.736	-.139
		L	-.694*	.245	.038	-1.355	-.033
	M	Q	.438*	.111	.004	.139	.736
		L	-.256	.156	.365	-.677	.165

Based on estimated marginal means

*. The mean difference is significant at the .05 level

a. Adjustment for multiple comparisons: Bonferroni.

15 Appendix 15

Shown is the raters' percent agreement in each of the subtests in Loudness experiment 2 (Chapter 5).

The data shows percentage of the raters that assigned ratings above +/- 1 to a particular voice quality stimulus. Ratings within the range of -1 to +1 are interpreted here as suggestive of 'no affect' signalled. Cases in which 75% of the raters or more agreed in voice to affect association are shown in bold type.

Table 15.1. *Raters' percent agreement in Loudness experiment 2. Cases of 75% agreement and above in voice to affect association are shown in bold type.*

Apologetic-Indignant			
	apologetic	indignant	no affect
whispery_Q	**100**	0	0
breathy_Q	**88**	13	0
lax-creaky_Q	**75**	13	13
modal_Q	38	56	6
tense_Q	6	**88**	6
whispery_M	**100**	0	0
breathy_M	**94**	6	0
lax-creaky_M	**81**	13	6
modal	6	**81**	13
tense_M	0	**100**	0
whispery_L	**81**	6	13
breathy_L	**75**	13	13
lax-creaky_L	**75**	19	6
modal_L	0	**94**	6
tense_L	0	**100**	0
Bored-Interested			
	bored	interested	no affect
whispery_Q	69	25	6
breathy_Q	69	25	6
lax-creaky_Q	**88**	13	6
modal_Q	44	38	19
tense_Q	6	**88**	6
whispery_M	50	31	19
breathy_M	63	38	0
lax-creaky_M	**94**	6	0
modal	19	63	19
tense_M	6	**94**	0
whispery_L	6	**75**	19
breathy_L	13	**75**	13
lax-creaky_L	**81**	13	6
modal_L	6	**94**	0
tense_L	6	**94**	0
Intimate-Formal			
	intimate	formal	no affect
whispery_Q	**100**	0	0
breathy_Q	**88**	6	6
lax-creaky_Q	**94**	6	0
modal_Q	25	56	19
tense_Q	6	**88**	6
whispery_M	**100**	0	0
breathy_M	**81**	6	13
lax-creaky_M	**94**	6	0
modal	0	**94**	6
tense_M	0	**100**	0
whispery_L	63	38	0
breathy_L	56	19	25
lax-creaky_L	**94**	6	0
modal_L	0	**100**	0
tense_L	0	**100**	0

Table 15.2. *Raters' percent agreement in Loudness experiment 2. Cases of 75% agreement and above in voice to affect association are shown in bold type.*

Relaxed-Stressed			
	relaxed	stressed	no affect
whispery_Q	**81**	13	6
breathy_Q	**88**	6	6
lax-creaky_Q	**94**	0	6
modal_Q	38	50	13
tense_Q	6	**94**	0
whispery_M	**81**	19	0
breathy_M	69	13	19
lax-creaky_M	**100**	0	0
modal	0	**75**	25
tense_M	0	**100**	0
whispery_L	31	63	6
breathy_L	44	56	0
lax-creaky_L	**81**	19	0
modal_L	6	**94**	0
tense_L	0	**100**	0
Sad-Happy			
	sad	happy	no affect
whispery_Q	**88**	13	0
breathy_Q	**81**	19	0
lax-creaky_Q	**88**	6	6
modal_Q	31	44	25
tense_Q	6	69	25
whispery_M	69	31	0
breathy_M	**75**	19	6
lax-creaky_M	**88**	6	6
modal	0	**81**	19
tense_M	13	**88**	0
whispery_L	44	56	0
breathy_L	56	44	0
lax-creaky_L	**88**	13	0
modal_L	6	**88**	6
tense_L	0	**94**	6
Scared-Fearless			
	scared	fearless	no affect
whispery_Q	**81**	13	0
breathy_Q	50	38	13
lax-creaky_Q	25	56	19
modal_Q	56	38	6
tense_Q	63	31	6
whispery_M	63	31	6
breathy_M	44	44	13
lax-creaky_M	38	56	6
modal	25	69	6
tense_M	44	50	6
whispery_L	63	25	13
breathy_L	38	63	0
lax-creaky_L	6	**94**	0
modal_L	31	63	6
tense_L	31	63	6

16 Appendix 16

Raters' agreement in Loudness experiment 2 (Chapter 5).

Pearson's *r* correlation coefficients were calculated showing correlation of each individual rater with the group mean for a particular subtest. This allows one to ascertain what raters within each group of listeners were found to assign ratings very different to those of a group as a whole.

These coefficients are presented in Table 16.1 below. Cases of Pearson's *r* below 0.5 (suggestive of substantial difference between individual raters and a group as a whole) are shown in bold type.

Table 16.1 *Pearson's* r *showing correlation between the ratings of individual raters and the group mean.*

	Pearson's r correlation with the group mean per test					
Subject	**Apologetic -Indignant**	**Bored-Interested**	**Intimate-Formal**	**Relaxed-Stressed**	**Sad-Happy**	**Scared-Fearless**
S_1	0.96	0.92	0.99	0.95	0.98	**0.49**
S_2	0.90	0.73	0.94	0.75	0.90	0.77
S_3	0.96	0.94	0.93	0.96	0.90	**0.00**
S_4	0.95	0.97	0.80	0.94	0.98	0.62
S_5	0.93	0.90	0.97	0.96	0.69	0.50
S_6	0.88	0.77	0.91	0.94	**0.36**	0.61
S_7	0.88	0.96	0.92	0.90	0.92	**0.34**
S_8	0.93	0.94	0.96	0.73	0.95	0.65
S_9	0.99	0.95	0.98	0.99	0.95	**0.40**
S_10	0.90	0.64	**0.41**	0.79	0.88	**0.37**
S_11	0.88	0.76	0.96	0.93	0.94	**0.00**
S_12	0.90	0.92	0.96	0.91	0.56	0.56
S_13	0.68	0.97	0.93	0.93	0.88	0.74
S_14	0.79	0.94	0.72	0.78	0.77	**0.40**
S_15	0.97	0.85	0.94	0.83	0.93	0.63
S_16	0.89	**-0.54**	0.95	0.97	**-0.46**	**0.22**

17 Appendix 17

Raters' agreement in Loudness experiment 2 (Chapter 5).

Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for each individual rater, for each loudness normalised stimulus (Series M) and its quieter (Series Q) and louder (Series L) versions, for each subtest. Each dot in the plots corresponds to the data for one rater. Group mean and standard deviation are shown as a pink dot.

The data allow one to ascertain the extent of agreement/disagreement the raters demonstrated in rating a particular stimulus.

The data also allow one to infer the degree of each rater's confidence in voice to affect association. Given the range of the rating scale, standard deviation of 1 and below (across 10 randomisations) suggests relatively confident voice to affect association.

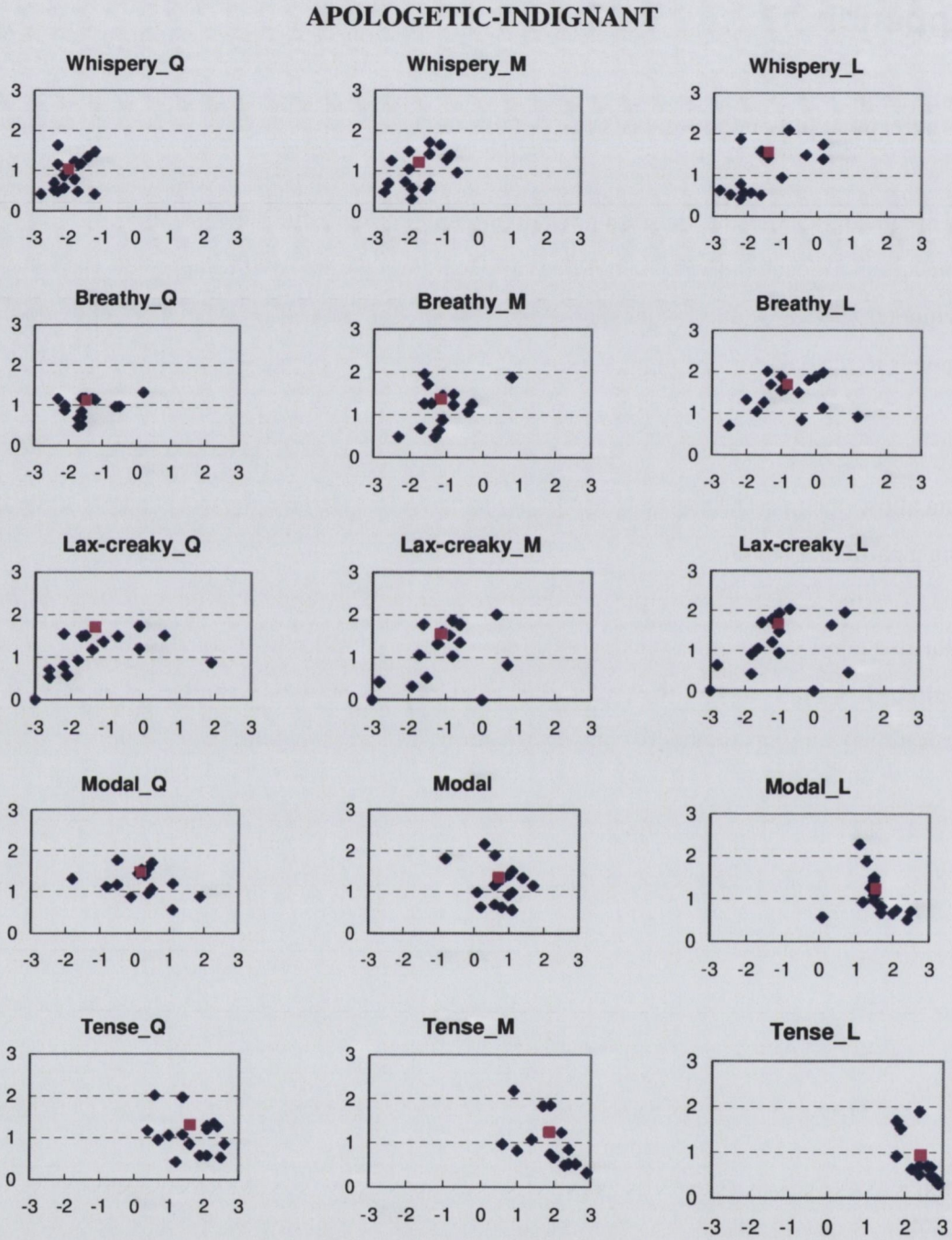


Figure 17.1. *Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for individual raters. Each dot in the plots corresponds to the data for one rater. Group mean and standard deviation are shown as a pink dot.*

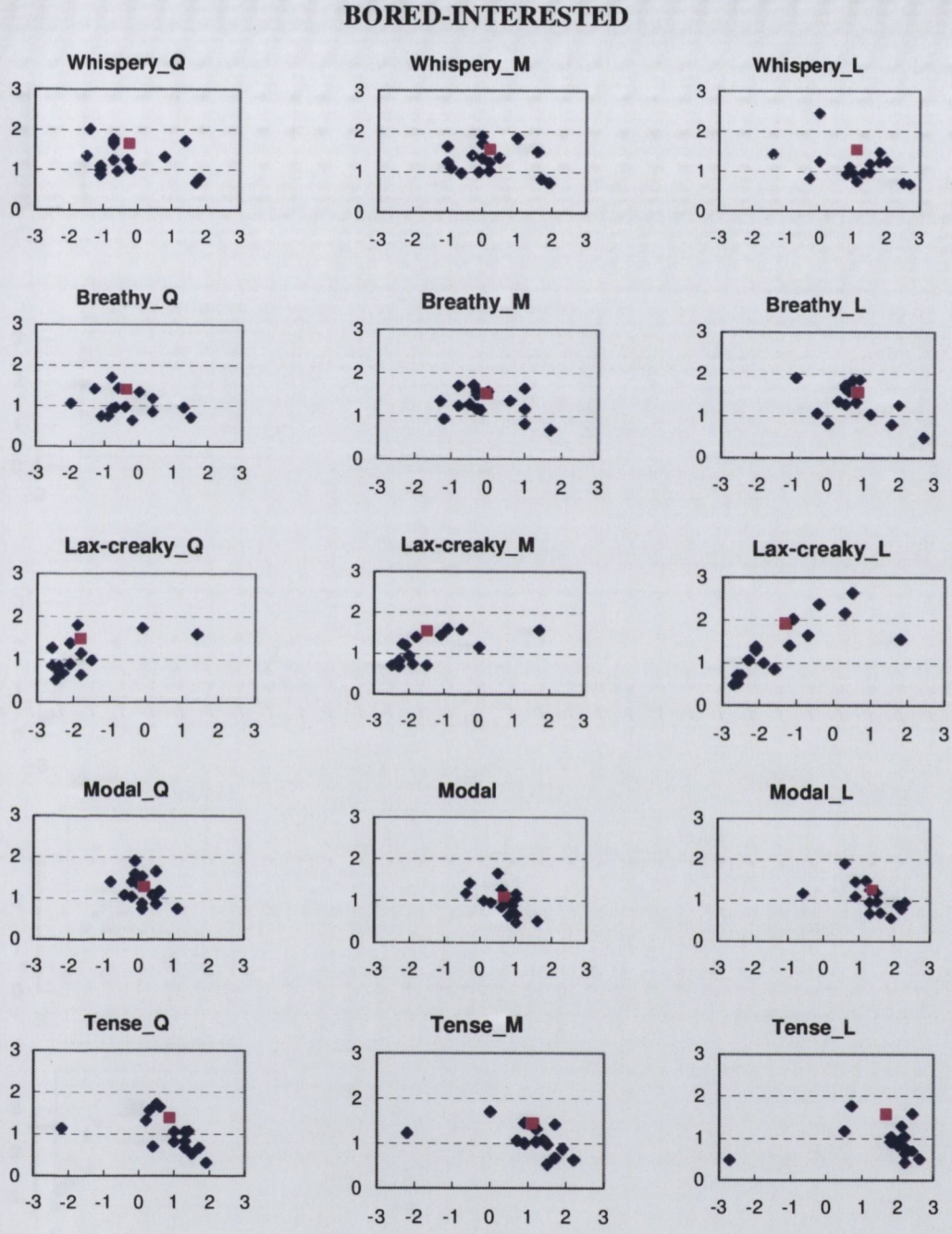


Figure 17.2. *Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for individual raters. Each dot in the plots corresponds to the data for one rater. Group mean and standard deviation are shown as a pink dot.*

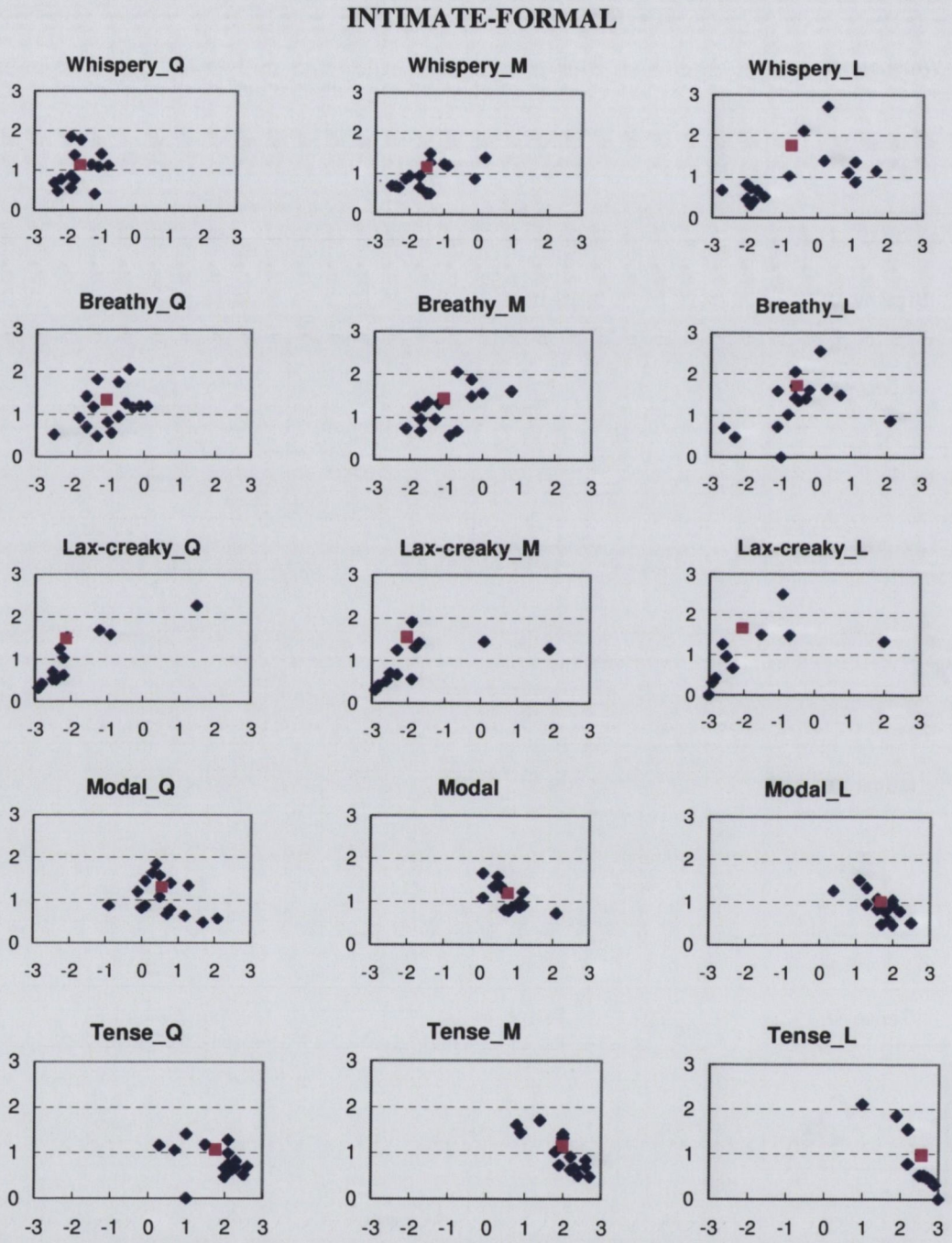


Figure 17.3. *Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for individual raters. Each dot in the plots corresponds to the data for one rater. Group mean and standard deviation are shown as a pink dot.*

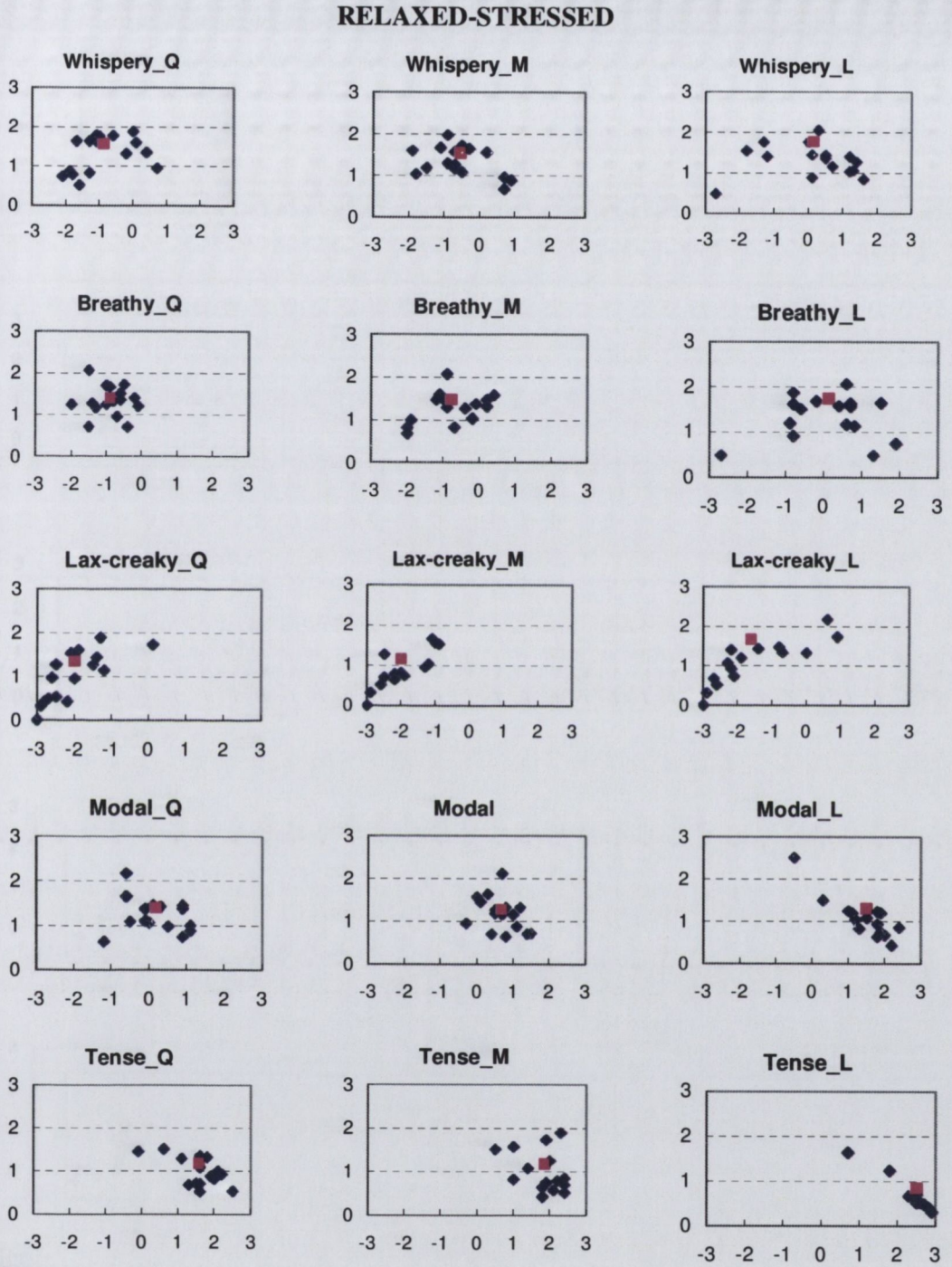


Figure 17.4. *Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for individual raters. Each dot in the plots corresponds to the data for one rater. Group mean and standard deviation are shown as a pink dot.*

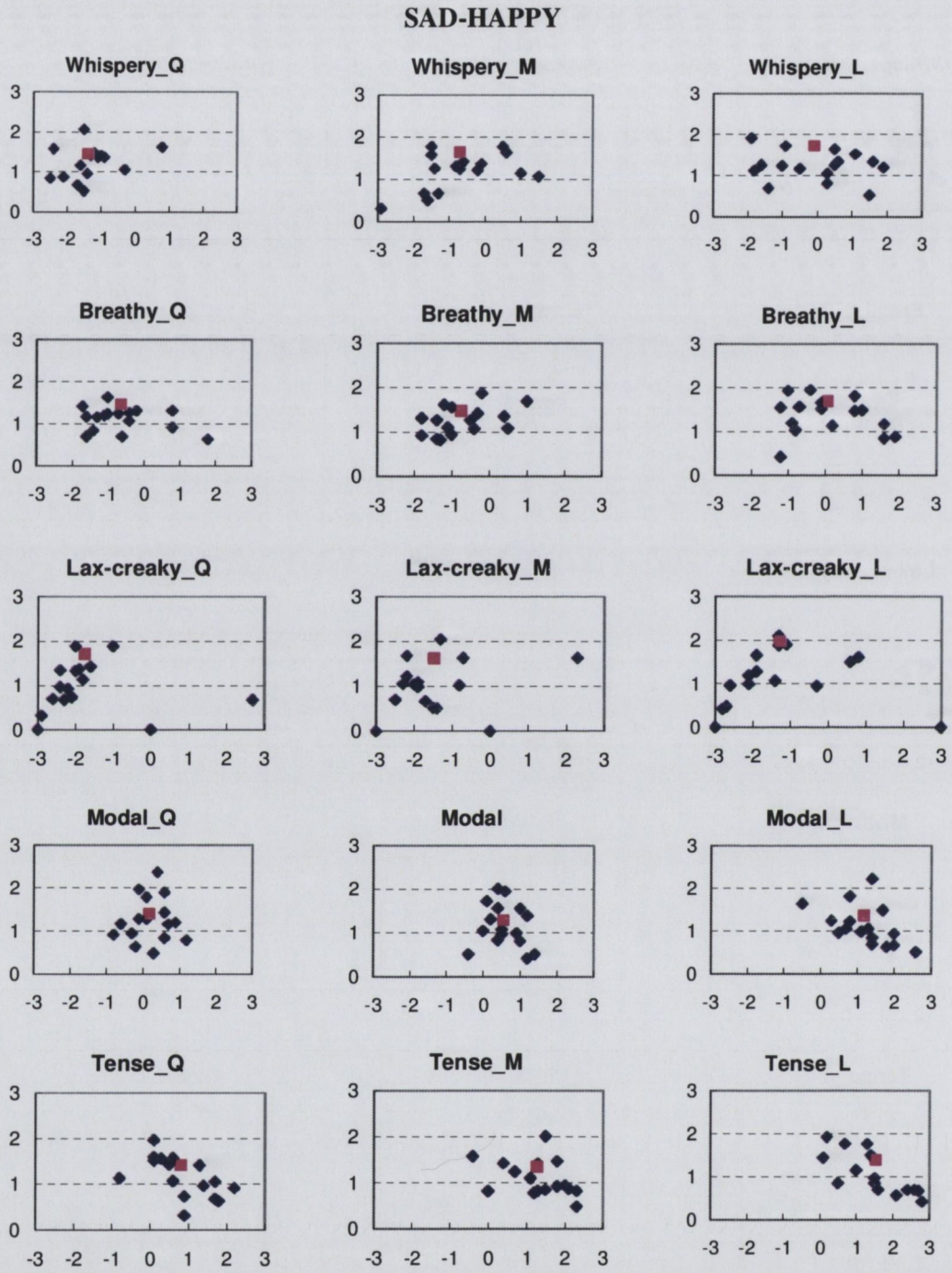


Figure 17.5. *Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for individual raters. Each dot in the plots corresponds to the data for one rater. Group mean and standard deviation are shown as a pink dot.*

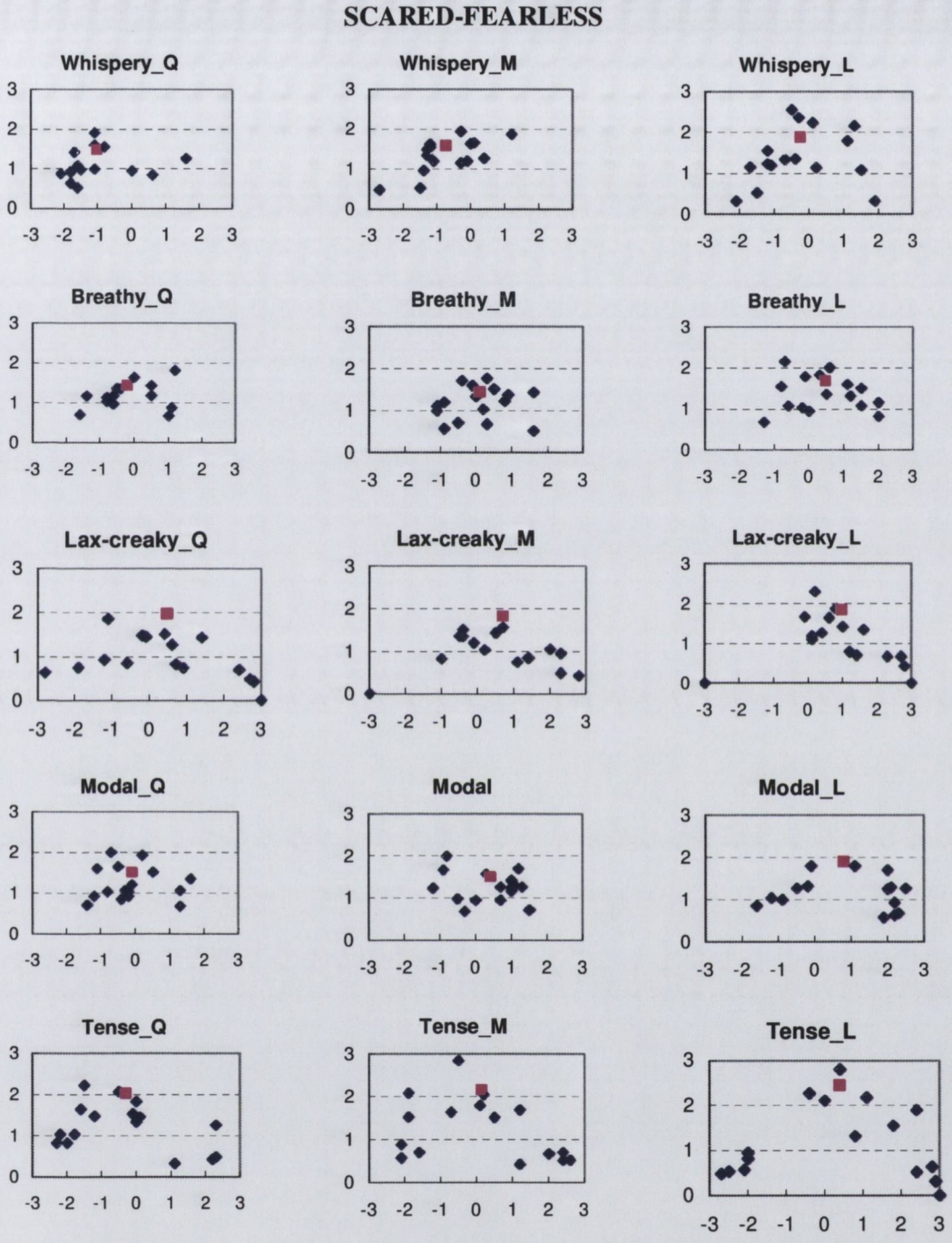


Figure 17.6. *Plotted are mean (x axis) and standard deviation (y axis) of the ratings across 10 randomisations for individual raters. Each dot in the plots corresponds to the data for one rater. Group mean and standard deviation are shown as a pink dot.*