Language matters in the perception of affect from voice

Irena Yanushevskaya, Ailbhe Ní Chasaide, Christer Gobl

Phonetics and Speech Laboratory, Centre for Language and Communication Studies, Trinity College Dublin, Ireland

yanushei@tcd.ie, anichsid@tcd.ie, cegobl@tcd.ie

Abstract

This paper presents the general results of a cross-language study of perception of affect from voice. A perception test using a range of synthesised voice quality stimuli was administered to four groups of subjects, speakers of Hiberno-English, Russian, Spanish and Japanese. The study aims to clarify how variations in voice quality (in synthesised stimuli) can evoke different affective colouring for subjects from 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. The results suggest both universal and language/culture-specific trends in voice to affect association.

Introduction

The human voice is a fascinating instrument, and its role in human interaction is fundamental. We are left in awe of 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 prone 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 importance of voice quality (the voice source characteristics related primarily to the mode of vibration of the vocal folds) in the vocal expression and communication of emotion is widely acknowledged, see, for example, (Scherer 2003). 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. 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 (Gobl and Ní Chasaide 2010). 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 (Gobl 2003; Ní Chasaide and Gobl 2004; Ní Chasaide, Yanushevskaya et al. 2011). 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.

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 and Scherer 1996; Scherer 2003; Bänziger and Scherer 2007), 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.

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. 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.

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 (2003), 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, and is used in this study.

The cross-language perception study presented here 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.

One of the main questions addressed in this cross-language study is whether the mapping of voice quality to affect is universal or language specific. In general, culture-specific influence on emotion expression/communication is opposed to universal biological factors (Zinken, Knoll et al. 2008; Ogarkova, Borgeaud et al. 2009). On the one hand, vocal expression of biologically based ('hardwired', 'basic') emotions represents a universal (often involuntary) behaviour related to distinct physiological changes such as muscle tension and sympathetic arousal (fight-or-flight response). These emotions tend to be 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. Here cultural factors govern the conventions of affect expression. Thus, the cognitive component is culturally informed.

This study aims at throwing light at what is universal and what may be culture-specific in the way subjects with different language/culture background perceive affective colouring from different voice qualities.

MATERIALS AND METHOD

Synthesised stimuli

The voice quality stimuli were based on prior analyses (Gobl and Ní Chasaide 2010) as well as on the broader literature on the acoustic characteristics of different voice qualities. The main objective of the study is to explore the mapping of voice quality to affect in each of the four language groups (Hiberno-English, Russian, Spanish and Japanese). The listeners' reactions were elicited not to individual dimensions of voice quality, such as spectral slope or breathiness, but rather to the holistic voice quality entity, such as breathy voice or tense voice. A short utterance [ˈjɑ aˈjø] was synthesized in 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 according to the framework in (Laver 1980). A detailed description of stimuli generation is given in (Gobl and Ní Chasaide 2003). The synthesized voice quality stimuli: whispery, breathy, lax-creaky, tense and modal were then combined with affect-related f_0 contours described in (Mozziconacci 1995) which varied in the magnitude of f_0 excursions and dynamics (f_0 contours fear, sadness, boredom, joy and indignation) in order to ascertain whether voice quality cues (particularly to strong emotions) become more effective when major f_0 perturbations are included. Furthermore, it was of interest 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. Thus, the material for the listening test included three groups of stimuli. The first of these, 'VQ only', included a range of distinct voice qualities. The second, 'f_0 only' group simply involved

manipulations to the 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. Thus, for example, whispery voice was combined with f_0 fear, breathy voice with f_0 sadness, lax-creaky with f_0 boredom etc. The stimuli could be further grouped into five 'Affect groups' each of which included three different types of stimuli manipulation, 'VQ only', 'VQ+ f_0' and 'f_0 only'. Thus, 'Affect group sad' included (i) breathy voice, (ii) breathy voice combined with f_0 sadness and (iii) f_0 sadness combined with modal voice (see Table 1 below).

Listening test

The listening test was conducted as outlined in (Gobl and Ní Chasaide 2003) as six subtests. In each subtest, the participants were asked to listen to the stimuli in random order and to assess the affective coloring of each stimulus on a 7-point rating scale with the polar opposite affective labels placed on each side of the scale. The scale allowed to rate the affective coloring of each stimulus as strong (+/-3), moderate (+/-2), mild (+/-1) or no affect (0). The pairs of affective labels were *apologetic-indignant, bored-interested, intimate-formal, sad-happy relaxed-stressed,* and *scared-fearless.* 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 (Juslin and Laukka 2003; Douglas-Cowie, Cowie et al. 2006; e.g., Baron-Cohen 2007). 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.

Table 1: Voice quality stimuli and affect-related f_0 contours used in the cross-language study.

'Affect group'	Voice quality stimuli	Affect–related f_0 contours
fear	whispery	f_0 fear
sadness	breathy	f_0 sadness
boredom	lax-creaky	f_0 boredom
joy	tense	f_0 joy
indignation	tense	f_0 indignation
neutral	modal	Neutral

Participants and translation of affective labels

The participants were speakers of Irish-English, Russian, Spanish and Japanese, 20-21 subjects in each language group; the groups were gender balanced. The selection of speakers was motivated by what is known about the use of voice quality and f_0 in each of these languages as well as by impressionistic observations. The participants were given instructions in their respective languages, and they were given an opportunity to ask for clarification as regards the test procedure prior to the test. The translation of the labels from English was undertaken by at least two native speakers of the respective languages who had a good command of English and who were also familiar with the nature and aim of the research and the purpose of the rating scales. The translators were asked to discuss the translation possibilities and to agree on the best possible choice of the labels in order to keep the translation accurate and to maintain the polarity of the affective labels.

Statistical analysis

A mixed type repeated measures ANOVA was conducted to assess the statistical significance of the differences in voice to affect associations across different language background groups. A complex mixed design 4x3x5 was used. The independent variables were Language (4), Stimulus Type (3), and Affect Group (5) while the dependent variable was the affective rating that each stimulus yielded. The interrater agreement was assessed using the single measures Intraclass Correlation Coefficient.

In the discussion of results we will primarily focus on ratings above +/-1 (above the grey area in Figs. 1-2). This threshold is admittedly arbitrary, but it allows us to focus on more robust and consistent voice to affect associations.

RESULTS AND DISCUSSION

All non-modal ‘VQ only’ stimuli get associated with a number of affects, although languages may vary in the choice of stimulus associated with a particular affect and in the strength of the affective rating accorded to a particular stimulus. 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 (Gobl and Ní Chasaide 2003; Campbell 2004), 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.

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.

Different f_0 contours available in the range showed different potency in affect signalling. Stimuli 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.

A few examples below will illustrate to what extent we find agreement among the groups of listeners with different language background and what distinct language-specific trends emerge.

Evidence of cross-language agreement

For a number of affective states, a considerable cross-language agreement was found in the way synthesised stimuli signal affect. There are of course some differences in the strength of affective ratings demonstrating disparity in the relative potency of the stimulus in cueing the affect for a particular language. This 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 the language groups tested, with differences showing up in the strength of the affective rating: whispery + f_0 'fear' signalling *apologetic* and *scared*, lax-creaky and lax-creaky + f_0 'boredom' signalling *bored* and *sad*; tense signalling *indignant* and *fearless*; tense + f_0 'indignation' and modal + f_0 'indignation' signalling *interested;* modal + f_0 'fear' signalling *scared*. This is illustrated in Fig. 1, for the *bored-interested* subtest, for a selection of stimuli.

Evidence of cross-language differences

In two subtests, *intimate-formal* and *relaxed-stressed*, the cross-language difference was substantial. The *intimate* 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 Irish-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 Irish-English and Russian, qualities that did not evoke *intimate* for the Span ish and Japanese listeners. There was a conspicuous gap in the cueing of *formal* with any of the stimuli presented here for the Japanese listeners (Fig. 2, panels A-B). There was no affective response from the Japanese in the *stressed-relaxed* test for any of the stimuli presented here (Fig. 2, panels C-D).

Figure 1: Languages agree: *bored-interested* (selection). Asterisks show statistically significant differences in stimuli ratings across language groups.

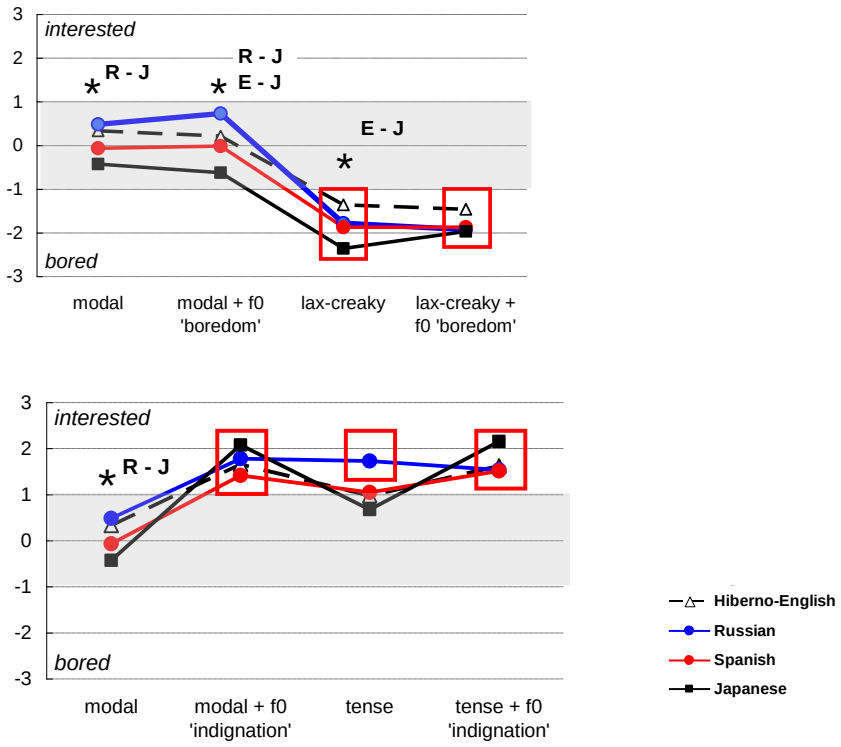
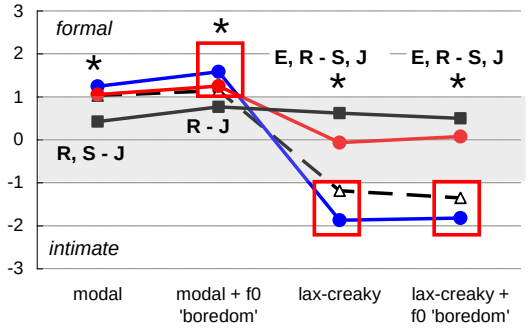
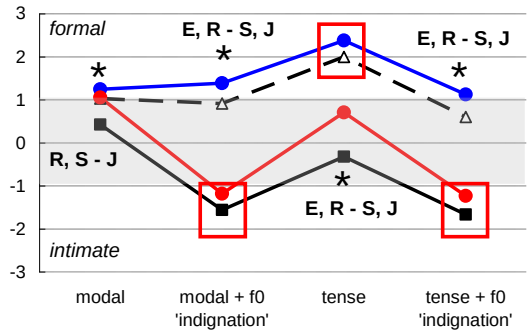


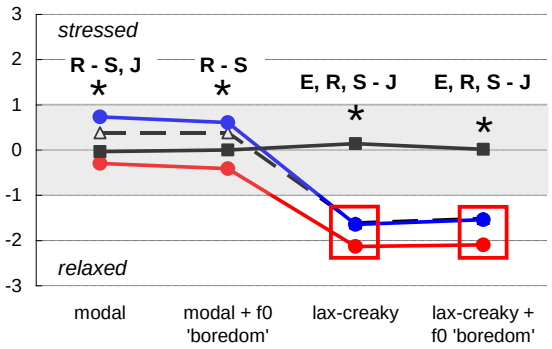
Figure 2: Languages disagree: *intimate-formal* (selection, panels A-B) and *relaxed-stressed* (selection, panels C-D). Asterisks show statistically significant differences in stimuli ratings across language groups.



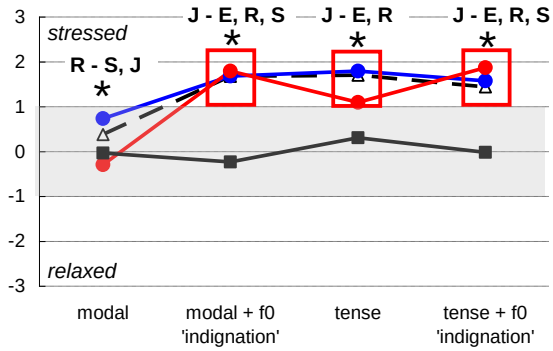
A



B



C



D

These differences are likely to be linked to underlying cross-cultural differences. 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. 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. E.g., 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. There is a possibility that the affective labels are not equivalent for the different languages. Intuitively, we did not feel that the semantic differences were likely to explain the very different voice to affect mappings emerging in this study. However, this issue does suggest interesting future research where one could try to elicit how stimuli such as these map to a richer range of affective states. Another factor that could contribute to this kind of cross-language differences would be the neutral voice (baseline) prevalent in these different languages. Impressionistically, 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. This is an interesting area in itself which intersects with how voice is used in a particular language to signal affect.

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 speaker of the European languages tested rely more heavily on voice quality cues.

CONCLUSIONS

The stimuli presented in this study are necessarily a very limited subset of the infinite variety that exists in real life. Certain voice qualities have been omitted (e.g., harsh voice and falsetto), and extreme versions of these qualities have not been included. As

for the f_0 contours, although a reasonable sampling of the types of differences in f_0 level and f_0 dynamics often described in the literature were included, these are also a limited selection of the endless possibilities that exist. The same points have to be made about the combined stimuli.

Where strong affective signalling emerged in the listening test, but with the different languages choosing very different voice qualities for such signalling (as in the case of *intimate* here), 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 listeners responses to the *stressed/relaxed* and to the *formal* affects, there is a gap in the affective signalling, we have to make rather different conclusions. It is 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. Production data, e.g., from spontaneous speech corpora (Iida, Campbell et al. 1998), which might prompt some more appropriate voice qualities (and f_0 contours) which could be tested within the kind of perception experiment presented here.

ACKNOWLEDGEMENTS

The research was supported by the EU Sixth Framework Network of Excellence HUMAINE; the Science Foundation Ireland (Grant 07 / CE / I 1142) as part of the Centre for Next Generation Localisation (www.cngl.ie); and Grant 09/IN.1/I2631 (FASTNET).

REFERENCES

Banse, R. and K. R. Scherer (1996). "Acoustic profiles in vocal emotion expression." Journal of Personality and Social Psychology **70**(3): 614-636.

Bänziger, T. and K. R. Scherer (2007). Using actor portrayals to systematically study multimodal emotion expression: the GEMEP corpus. Affective Computing and Intelligent Interaction (ACII 2007). A. Paiva and R. W. Picard. Berlin, Springer-Verlag. **4738:** 476-487.

Baron-Cohen, S. (2007). Mind Reading: the Interactive Guide to Emotions. London, Jessica Kingsley Publishers.

Campbell, N. (2004). Perception of affect in speech - towards an automatic processing of paralinguistic information in spoken conversation. Interspeech 2004 - ICSLP, Jeju Island, Korea.

Douglas-Cowie, E., R. Cowie, et al. (2006). HUMAINE Human Machine Interaction network on Emotions Mid Term Report on Database Exemplar Progress.

Gobl, C. (2003). The Voice Source in Speech Communication. Doctoral thesis. Stockholm, KTH, Department of Speech, Music and Hearing.

Gobl, C. and A. Ní Chasaide (2003). Amplitude-based source parameters for measuring voice quality. VOQUAL'03, Geneva, Switzerland.

Gobl, C. and A. Ní Chasaide (2003). "The role of voice quality in communicating emotion, mood and attitude." Speech Communication **40**(1-2): 189-212.

Gobl, C. and A. Ní Chasaide (2010). Voice source variation and its communicative functions. The Handbook of Phonetic Sciences. W. J. Hardcastle, J. Laver and F. E. Gibbon, Blackwell Publishing Ltd**:** 378-423.

Iida, A., N. Campbell, et al. (1998). Acoustic nature and perceptual testing of corpora of emotional speech. 5th International Conference on Spoken Language Processing, Sydney, Australia.

Juslin, P. N. and P. Laukka (2003). "Communication of emotions in vocal expression and music performance: different channels, same code?" Psychological Bulletin **5**: 770-814.

Laver, J. (1980). The Phonetic Description of Voice Quality. Cambridge, Cambridge University Press.

Mozziconacci, S. (1995). Pitch variations and emotions in speech. XIIIth International Congress of Phonetic Sciences, Stockholm.

Ní Chasaide, A. and C. Gobl (2004). Decomposing linguistic and affective components of phonatory quality. Interspeech 2004, Jeju Island, Korea.

Ní Chasaide, A., I. Yanushevskaya, et al. (2011). Voice source dynamics in intonation. XVII International Congress of Phonetic Sciences, Hong Kong, China.

Ogarkova, A., P. Borgeaud, et al. (2009). "Language and culture in emotion research: a multidisciplinary perspective." Social Science Information **48**(3): 339-357.

Scherer, K. R. (2003). "Vocal communication of emotion: a review of research paradigms." Speech Communication **40**: 227-256.

Zinken, J., M. Knoll, et al. (2008). Universality and diversity in the vocalisation of emotions. Emotions in the Human Voice. K. Izdebski. San Diego, CA, Plural Publishing**:** 185-202.