

Vocal communication of emotion:
Exploring the contribution of local and global shifts to the voice
source and the impact of the visual context

Thesis submitted for the degree of Doctor of Philosophy

2025

By

Anna Maria Giovannini

Supervisor: Professor Christer Gobl

Phonetics and Speech Laboratory

School of Linguistic, Speech and Communication Sciences

Trinity College Dublin, the University of Dublin

DECLARATION

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work. I confirm that the work has been completed in full compliance with relevant university policies, including the Academic Integrity policy, the Trinity Policy on Good Research Practice, and the guidance on AI and GenAI in Teaching, Learning, Assessment and Research.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

Signature of the Author: _____

A handwritten signature in black ink, appearing to read 'Anna Maria Giovannini', is written over a horizontal line. The signature is stylized with a large initial 'A' and a long, flowing tail.

Anna Maria Giovannini

12 September 2025

‘Courage need not be remembered, for it is never forgotten.’

– The Legend of Zelda

SUMMARY

This thesis explores aspects of the mapping of voice quality to affect and the impact of the visual context in emotion perception. The first phase of the study involves a thorough analysis of recorded acted data through manual inverse filtering and model matching to the Liljencrants-Fant model, used as a basis to create contour visualizations and estimate the changes to the voice source related to emotion expression. The recorded stimuli consisted of Irish sentences spoken in *happy*, *angry*, *sad*, *bored* and *relaxed* affects. This is followed by the development of a novel methodology on different ways of applying acoustic modifications to the voice source parameters f_0 , EE and RD to a neutral synthesized baseline – namely in terms of global (utterance-wide) or local (aligned to the prosodic structure of the sentence, i.e., pitch accents) changes.

Global changes were done by shifting parameter values across the entire utterance, based on regression analysis values extracted from each recorded stimulus of a particular affect. Local changes were done using an anchor-point approach aligned to the to the prosodic structure of the utterances, modifying only the parameter values within pitch accents relative to the global regression analysis. Finally, combined changes were made by bringing global and local shifts together. These changes were then applied to the neutral baseline using the Voice Source Generator, along with a final set of adjustments needed to ensure high quality audio with no artefacts and presented to groups of listener judges to evaluate the effectiveness of the modifications system.

Four main studies were conducted. The first two focus on the vocal stimuli and aimed to investigate whether global, local or combined changes were more effective in cueing the intended affect, and whether individual affects were well differentiated (especially within activation categories). Experiment 1 applied the emotional shifts to a recorded neutral baseline, presented to listeners in a multiple-choice task that included all affects produced and an *interested* option. The results of the experiment were mixed. The *happy* transformations and the *angry* combined stimulus were the most successful in terms of identification, but the low activation affects were only identified in terms of their activation category. The types of transformations also varied in effectiveness for each affect

- the local had significant effects for *happy*, the combined was the most successful for *angry*, and the global performed best for *sad*.

Experiment 2 used a neutral baseline produced by an Irish TTS voice featuring the same speaker. In this test, listeners' judgements were elicited in terms of three scales of opposing affects: *sad-happy*, *relaxed-angry*, and *bored-interested*. Listeners judged whether and how strongly a particular emotion was perceived. Here, the combined stimuli were the most effective, particularly for *happy* and *relaxed*, followed by the global stimuli for the other affects. While there seemed to be some additive effect of the global and local stimuli, the local stimuli performed much more poorly. It is speculated that this was due to the baseline change. The low activation stimuli were poorly differentiated, while results for high activation states were mixed – only *happy* was uniquely cued from *angry*, but both affects were often also perceived as *interested*.

The second set of experiments investigated the effect of the visual context on the perception of emotion, using the combined stimuli from experiment 2. A set of 3D animated facial expressions were created for each affect (with the addition of *interested* as an opposing visual cue to the *bored* affect). Listeners were asked to judge the emotion perceived for the following sets of stimuli: voice-only; voice with congruent visuals; and voice with incongruent visuals. In experiment 3, incongruent stimuli were created by combining voices and faces of opposing affects (*sad + happy*, *relaxed + angry*, *bored + interested* and vice-versa). For experiment 4, the incongruent stimuli were created by swapping voices and faces within activation categories (e.g., *angry + happy*, *sad + bored*). The same rating scales from experiment 2 were used.

Results from this set of experiments indicate that, while the addition of the facial expressions did not improve affect identification rates, it did improve differentiation between affects to a degree – particularly by separating affects of opposite valence within the same activation category. The incongruent stimuli did seem to have an effect the perception of the emotion in the voice, although for the most part the voice was still dominant. This may indicate that different affects rely more on different information channels to be successfully identified. The shifts and methodology proposed here were relatively successful, especially across affect activation categories, indicating potential success for implementation in terms of emotion expression.

ACKNOWLEDGEMENTS

This thesis would not have been possible without the support and encouragement of many. First, I would like to thank my supervisor, Prof. Christer Gobl, for all the support, advice, kindness, patience and understanding throughout this project. Another big thank you goes to Prof. Ailbhe Ní Chasaide for the advice on Irish and emotional research, and the help in finding informants for the experiments. Thank you also to Prof. Neasa Ní Chiaráin and her family for the support and kindness extended to me as I found a place in Ireland. I would also like to thank my examiners, Dr. Irena Yanushevskaya and Prof. Mattias Heldner, for the very helpful feedback and insight, patience and kindness.

My utmost gratitude also goes to the Irish Research Council for funding the most crucial years of this research. I'd also like to thank the HEA for the additional funding that allowed me to finish my project with some additional peace of mind.

A big thank you goes to my colleagues at the Phonetic and Speech Laboratory at Trinity College Dublin. Special thanks go to Zihan Wang for the friendship, company and all the help through this project, from literally developing the software that allowed it to happen to all the times we worked on something together. I'd also like to thank Dr. Maria O'Reilly for the help with the intonation analysis and the great company, and Dr. Daniel Dempsey for the invaluable help with the statistical analyses. To everyone who's been in the lab over the last four and a half years – thank you for your friendship, for all the good times we had and for your support through this journey. You guys are all champions.

A huge thank you also goes to my mom and dad, Marcia and Fabrizio Giovannini, for all the love, support, for believing in me through these most challenging times and for always being there when I needed most. I love you both with all my heart. Thank you also to our cat Lena for all the joy she brings whenever I visit home, and to all my family for always being so kind and supportive.

Finally, a huge thank you to all my friends, old and new, from back home and all over the world. I am so incredibly grateful to all of you for your patience, encouragement, love, hugs, D&D adventures, and for being there and believing in me more than I ever could. You have no idea how much it all meant, and I could never have gotten here without you reminding me of all the great things out there.

TABLE OF CONTENTS

1.	Introduction	1
1.1.	Research Context and Motivation.....	2
1.1.1.	Expression of Emotion in Speech Recognition and Synthesis.....	2
1.1.2.	The Context of Irish	4
1.2.	Research aims.....	5
2.	Literature Review	8
2.1.	Theoretical Framework	8
2.1.1.	Emotion	8
2.1.1.3.	<i>Models for Emotion Research</i>	12
2.1.2.	Basics of Speech Production	18
2.1.3.	Acoustic Theory of Speech Production	26
2.1.4.	Voice Source Analysis.....	28
2.1.5.	Glottal Source Models.....	32
2.1.6.	Intonation and Prosodic Structures	41
2.1.7.	Irish Prosody and Intonation	45
2.2.	Research on the Acoustics of Affect.....	47
2.2.1.	Research Models on Vocal Expression of Affect	47
2.2.2.	Speech Data Acquisition.....	50
2.2.3.	Experimental Perception Approaches.....	53
2.2.4.	Acoustic Correlates of Emotion.....	57
2.2.5.	The Voice Source and Prosodic Structures.....	65
2.3.	The Role of Context.....	67
2.3.1.	Emotion as Multimodal	68
2.3.2.	The Visual Channel	70
2.3.3.	The Semantic Context	73
2.3.4.	The Cultural Context	74
2.3.5.	Individual Speaker Characteristics	76
2.4.	Affective Speech Synthesis.....	78
2.4.1.	Direct Emotional Speech Synthesis	78
2.4.2.	Emotional Voice Conversion	82
2.5.	Summary	89
3.	Methodology	93
3.1.	Production Data Acquisition	93
3.2.	Data Annotation and Segmentation	96
3.3.	Data Analysis and Visualization Methods	99

3.3.1.	Initial Data Analysis Approaches and Considerations	99
3.3.2.	Manual Analysis Methods	100
3.4.	Parameter Modification and Resynthesis Methods.....	102
3.5.	Visual Stimuli Creation Methods	105
3.6.	Perception Tests	106
3.6.1.	Discrimination-based Task.....	106
3.6.2.	Scale-based Task.....	107
3.7.	Statistical Analyses	108
4.	THE AFFECT MODIFICATION system	111
4.1.	Production Data Analysis.....	111
4.2.	Parameter Modification and Resynthesis Methods.....	119
4.2.1.	The Parameter Modification System.....	119
4.2.2.	Application of Parameter Shifts and Resynthesis through the VSG System	121
4.2.3.	Initial attempts and considerations.....	124
4.2.4.	Global Manipulations	128
4.2.5.	Local Manipulations	130
4.2.6.	Combined Manipulations	133
4.2.7.	Other transformation details	135
5.	Perception Experiment 1 – Recorded Neutral Baseline	138
5.1.	Modification Details	138
5.2.	The Happy Affect Pre-Test.....	149
5.2.1.	Final Manipulation Details.....	151
5.3.	Perception Tests	153
5.4.	Statistical Analysis and Results.....	154
5.5.	Discussion and Conclusions.....	167
6.	Perception Experiment 2 – ABAIR Baseline.....	170
6.1.	The ABAIR Baseline Neutral and Modification Details	170
6.2.	Perception Tests	183
6.3.	Statistical Analysis and Results.....	185
6.4.	Discussion and Conclusions.....	200
7.	Perception ExperimentS 3 and 4 – The Visual Channel.....	202
7.1.	Materials.....	203
7.1.1.	Vocal Stimuli	203
7.1.2.	Visual Stimuli	204
7.1.3.	Visual Stimuli Validation Experiment	216
7.2.	Experiment 3 – Visual Context with Cross-Activation Congruence.....	221

7.2.1.	Perception Test Design and Subjects	222
7.2.2.	Statistical Analysis and Results.....	223
7.2.3.	Discussion and Conclusions.....	238
7.3.	Experiment 4 – Visual Context with Within-Activation Congruence	240
7.3.1.	Perception Test Design and Subjects	241
7.3.2.	Statistical Analysis and Results.....	241
7.3.3.	Discussion and Conclusions.....	256
8.	Conclusions	258
8.1.	General Discussion and Conclusions	258
8.2.	Limitations of the Study	262
8.3.	Future Directions.....	264
	References.....	266
	Appendices	305
	Appendix A	305
	Appendix B	310
	Appendix C	333
	Appendix D	356
	Appendix E.....	379

1. INTRODUCTION

When we speak, we not only convey meaning through the words and sentences we utter, but also through the way in which we utter them. An important element of spoken human communication is the expression of emotion through tone of voice, indicating our response to the environment within certain social constraints. Knowing about the speaker's emotional state also allows listeners to infer their intentions and react accordingly (Buck, 1984; Juslin & Laukka, 2003; Plutchik, 1994). This aspect of human spoken communication is crucial to the development of social relationships (Ekman, 1992b).

It is likely that our ancestors utilized non-verbal vocal communication for most of our evolutionary history before speech and language as we know it today was developed (Scherer, 1982). Communication of emotion is considered by many researchers as 'the foundation of the social order in animals and humans' (Juslin & Laukka, 2003, p. 770), reflecting the environment in which they developed and being a key factor in determining the social structure of a species.

As such, both spoken and non-verbal expression of emotion are believed to have developed within certain environments, or contexts. It is not only limited by certain physical constraints (i.e., what the vocal apparatus is physically capable of producing) but also by ecological constraints, which relates to our degree of social cooperation as a species (Juslin & Scherer, 2005). While developments in human technology have allowed us to communicate across distances, it is widely accepted that human expression of emotion originally evolved through face-to-face interaction within inter-group contexts, by sharing spaces and developing societies and social identities through millennia. This may indicate that, while there has been substantial evidence that humans can perceive emotions only through vocal communication, context cues such as facial expressions and shared culture between speaker and listener still play a significant role in how such emotions are identified, perceived, and reacted to in day-to-day interaction.

The main aim of the present study is to explore aspects of the human expression of emotion within the context of the Irish language, and considering the eventual possibility of being able to manipulate synthetic speech through very specific changes to fundamental

dimensions of the voice towards certain target affects (i.e., emotions). The study investigates whether these voice manipulations are most effective when carried out in terms of global shifts (i.e., changes applied uniformly across the entire utterance), or in terms of local shifts that depend on the prosodic structure of an utterance (i.e., aligned to the pitch accented syllables of the utterance). Based on the analysis of recorded portrayals of emotion, a relatively simple system of modifications is applied to a neutral-non-emotional synthesized sentence, then presented to listener judges.

This study also explores the effects of the visual context on the perception of emotion to evaluate the effectiveness of these voice changes when more emotional information is present. This is prompted by past studies where it was found that an utterance with a given voice quality might evoke responses to different low or high activation states (e.g., Gobl & Ní Chasaide, 2003b; Murphy, Yanushevskaya, Ní Chasaide, & Gobl, 2022). It has been suggested that the finer differentiation among high or low activation states might depend on the context (visual or pragmatic) of the interaction.

1.1. Research Context and Motivation

1.1.1. Expression of Emotion in Speech Recognition and Synthesis

In the past few years, the field of speech science has seen increasing interest in understanding the non-verbal aspect of human speech for applications involving speech recognition and synthesis. In the medical field, speech analysis has been used in detecting disorders like Parkinson's Disease (Narendra, Schuller, & Alku, 2021) and clinical depression (Dubagunta, Vlasenko, & Magimai-Doss, 2019; Moore, Clements, Peifer, & Weisser, 2008). A review by Ding and Zhang (2023) even explores evidence of prosody profiles for patients with disorders such as Alzheimer's, schizophrenia, autism and depression, advocating for approaches to the study of context-driven comprehension and expression of pragmatic-affective prosody in order to shine a light on the communicative problems experienced by those with such disorders in terms of recognition and expression of emotion.

Speech synthesis has also been increasingly used to assist those with communicative disabilities and speech impairments. As the quality of speech synthesis improves, a wider variety of voices has been developed to the point that these synthetic voices may even be personalized to better match the speaker's own original voice before loss (Triandafilidi,

Tatarnikova, & Poponin, 2022), or something they perceive as appropriate for themselves (Creer, Cunningham, Green, & Yamagishi, 2013). The technology has also benefited those with visual impairments through screen readers (Ní Chasaide et al., 2017) and in education (Matoušek et al., 2020). It is therefore important that such speech synthesis tools are as natural and expressive as possible, given that they are not only tools for the integration (or reintegration) of those with speech disabilities into society (Mertl, Žáčková, & Řepová, 2018), but also given the evidence that listeners tend to retain more information from natural speech than from monotone, artificial synthesis (Hjelmquist, Dahlstrand, & Hedelin, 1992).

Speech recognition and synthesis has also recently seen some degree of application in language learning and education, specifically in the field of Computer-Assisted Language Learning (CALL). Recognition can help students with speaking practice, providing feedback on pronunciation, while synthesis can help those with visual impairments and provide pronunciation examples. In recent years, educational frameworks focused on learner context have also emerged, distinguishing between classroom and independent learning environments (e.g., Ní Chiaráin & Ní Chasaide, 2015). Such frameworks often explore the topic of learner motivation, recognizing that the difficulty of learning a new language can often drive students to frustration if they fail to utilize the language to their satisfaction (Dehghanzadeh, Fardanesh, Hatami, Talaei, & Noroozi, 2019). One reason for such failure is often the lack of regular practice and exposure to the target language, which is paramount to successful language acquisition (Hao, Liang, Wang, Liu, & Chen, 2021; Richter, 2019; Strawbridge, Sonesson, & Griffith, 2019). This is particularly critical for endangered languages with a limited presence in day-to-day life and a small number of speakers, which may particularly benefit from speech recognition and synthesis applications as supplemental practice.

It is therefore important for language learning applications to be engaging and motivating for the learner, for which the potential of immersive online environments has been widely recognized. Learning applications utilizing speech recognition and synthesis in their practice drills could improve the immersion aspect through natural, expressive speech synthesis. While a study by Young and Shishido (2022) showed that speech synthesis was less favored in second language listening activities by students due to its monotony despite

being perceived as easier to understand than natural speech, there is still evidence that speech synthesis provides good support for teachers and are well received by students in terms of its intelligibility (Matoušek et al., 2020), especially for the visually impaired. This desire to improve expressive synthesis for Irish is therefore part of the motivation for the present investigation, given its limited number of speakers and practice opportunities.

1.1.2. The Context of Irish

Irish is a Goidelic language of the Insular Celtic branch of the Celtic language family, indigenous to the island of Ireland. It is recognized as the first official language of the Republic of Ireland, and since 2022 as an official language of the European Union. However, centuries of colonization led to the decline of the language, something that was accelerated as a result of several factors in the nineteenth century such as the implementation of British imperialistic policies (McDermott, 2011), the great famine and the subsequent extreme levels of emigration. Today, the language is spoken as a community language only in the *Gaeltacht* areas, mostly along the western seaboard. Even in these areas, it is estimated that only 24% of the population speaks it as their first/home language (Ní Chasaide et al., 2017), and it was designated an endangered language by UNESCO (UNESCO, 2010).

Still, Irish is an obligatory subject up to school-leaving age in Ireland, and there are thus many L2 (second language) speakers. The teaching of Irish to L1 (first language) English-speaking children, however, presents numerous challenges, such as a lack of exposure to native speaker models of the language, a lack of resources, and particularly of modern motivating materials such as those available for other languages. To address these challenges, speech technologies (synthesis and recognition) have been developed and are being integrated into learning applications by the ABAIR research group at Trinity College Dublin (www.abair.ie). There is also an emphasis on developing applications that enable people with disabilities to engage with the language. One recent development has been an Irish language AAC (alternative and augmentative communication) system, which allows non-speaking people to communicate by selecting pictures/words which are concatenated to form sentences that are then spoken out by the synthetic voice (Barnes et al., 2023).

In such educational and disability applications, it would be highly desirable to be able to manipulate the synthetic voice so that the spoken message can convey the intended

emotion of the user accompanying the words and sentences. Towards such an eventual goal, the present dissertation explores some ways in which this might be achieved.

1.2. Research aims

This research project aims to explore how specific types of manipulation of vocal parameters may evoke emotion, and how the visual context impacts the emotion perceived by the listener.

The first phase of this work involves analytic and perceptual studies based on recordings of acted emotions, investigating the acoustic correlates of a set of affective states in terms of global and local changes in the acoustic parameters related to the voice source. This phase includes (1) the acquisition of emotion portrayals and voice source analysis of selected data; (2) the development of a method for voice source modifications to add emotional coloring to the output of a speech synthesizer (the precise modifications being guided by the production data and by existing literature on the topic); (3) the creation of expressive speech stimuli from a neutral synthetic baseline; and (4) the presentation of perception tests to naïve listeners to evaluate the effectiveness of the proposed system for emotional coloring.

This phase aims to provide insight into this project's main research question regarding how voice source changes to an utterance may signal emotion. Are emotions effectively conveyed through (i) global, utterance-wide shifts in voice source parameters across the entire utterance; (ii) through local shifts aligned to the prosodic structure of the sentence (i.e., shifts on the pitch accented syllables), or (iii) through a combination of global and local shifts?

The initial hypothesis is that the most effective emotional coloring will be global and local shifts combined, with the supposition that these effects are additive. This hypothesis comes from the reasoning that, because the combined shifts have the highest degree of deviation from the neutral baseline in terms of parameter values (compared to the global and local shifts separately) and therefore potentially the largest amount of affective information, they would be the most successful in cueing the target affect to listeners. This is also based on previous studies that indicate that there is no one-to-one mapping between

acoustic parameters and affect in global terms (Gobl & Ní Chasaide, 2010), and that affect signaling entails a combination of both utterance-wide shifts in voice parameters alongside local changes sensitive to prosodic structure (Murphy et al., 2022). A secondary hypothesis is that, between the local and global effects, the local shifts would confer better cueing of emotion, given their interactions with the prosodic structure of sentences (Arvaniti, 2020; Larrouy-Maestri, Poeppel, & Pell, 2024; Ní Chasaide & Gobl, 2004a), the evidence for the role of parameter contours in cueing emotion (Audibert & Aubergé, 2007; O'Reilly & Ní Chasaide, 2007), and findings that report global parameter levels as insufficient for emotion discrimination (Scherer, 1981).

The second phase of this study investigates the effect of the visual context on the perception of emotion, exploring how vocal expression of affect may interact with visual information. The inclusion of the visual stimuli in the present research is motivated by the future intended applications of the findings from this study. Because vocal manipulations are kept simple and may not successfully cue certain emotions, facial expressions were considered as a secondary major channel of emotional information that could be presented by users of the TTS systems during emotion expression. Therefore, the effectiveness of the proposed voice modifications was tested both for the voice only stimuli as well as with an added visual context in order to examine if the target affects explored in this study would be better cued or differentiated with the addition of the visual cues expected to be inherently present in future applications.

While the effects of added context to the vocal stimuli are well documented (one may instinctively assume that additional contextual information results in higher emotion recognition accuracy), this experiment was also conducted to investigate the contribution of each channel by combining vocal and visual expressions of emotion in both congruent (the same target emotion for both kinds of stimuli) and incongruent (a vocal stimulus targeting one emotion combined with a visual stimulus targeting a different emotion) manners. The congruent combinations follow the rationale that adding visual cues to the vocal stimuli would not only strengthen the cueing of the target affect (Scherer, 1986; Scherer & Scherer, 2011) but also aims to investigate whether a particular information channel is more effective in cueing particular emotions than the other (Juslin, Laukka, Harmat, & Ovsianikow, 2021; Scherer, 2003), especially those that are poorly

differentiated from the voice alone. On the other hand, the combination of incongruent stimuli investigates channel dominance when the vocal and visual cues conflict, exploring if and how the addition of the visual channel could sway listener judgements away from the target vocal affect (and vice-versa) (e.g., Collignon et al., 2008; Rigoulot & Pell, 2014).

Thus, the initial hypothesis is that the perception and interpretation of vocal stimuli is influenced by the visual stimuli with which it is presented – when congruent, the visual information strengthens the recognition of the target emotion; when incongruent, one of the channels (vocal or visual) is dominant in the listener’s perception of emotion. A secondary hypothesis is that congruent visual information enhances the differentiation of the perceived emotion among low or high activation groups of affects, which did not emerge as clearly differentiated in past studies (e.g., Gobl & Ní Chasaide, 2003b; Murphy et al., 2022; Ní Chasaide & Gobl, 2005; Yanushevskaya, Gobl, & Ní Chasaide, 2018) where voice qualities were mapped to emotion.

The methodology and findings of this research are intended as a proof of concept to, in the future, be applied to an Irish synthesizer and some of its current applications. Although not explored in this project, it is further hypothesized that the local changes in emotion signaling parameters within structurally distinct aspects of the phrase need to be calibrated to the baseline voice settings of the speaker. It is hoped that the methodology hereby presented is appropriate for use on any type of vocal profile, but also suggested that parameters be further tested and refined for different voices in future research.

The remainder of this thesis is structured as follows. A review of the relevant literature is presented in chapter 2, including an exploration of theories of emotion, voice source models and parameters used in this research, prosody, the role of context in how emotions are expressed and perceived, and developments in affective speech synthesis and voice manipulation. Chapter 3 details the methodology of the study. Chapter 4 details the results of the analysis of production data, the development of the system of voice modifications and how it is applied to the neutral baseline. Chapters 5 and 6 present the first set of experiments. Chapter 7 presents the visual context experiments, also detailing the creation of the visual stimuli. Finally, chapter 8 presents the overall conclusions of the project, bringing together all tests conducted for this thesis and detailing the limitations of the study and future directions for research.

2. LITERATURE REVIEW

This chapter presents a review of the literature in all areas relevant for the present research. Section 2.1 covers the theoretical framework of this thesis, starting with the definition of emotion as a concept, theories on the nature and models of emotion categorization for research, onto basic concepts of speech production and critical frameworks for describing voice quality and parameterizing the voice source for the proposed research, along with a summary on prosody and intonation.

Section 2.2 brings together emotion and speech production through a review of the current body of research on the acoustics of emotional speech, including research models on vocal expression of affect, methods of data collection and experimental perceptual approaches, the acoustic correlates of emotion from a wide range of studies and, finally, research on the relationship between the voice source and the prosodic structure of utterances.

Section 2.3 explores the role of context in the vocal expression and perception of emotion, starting with general studies on the multimodality of emotion and going into the effects of the visual channel and the semantic context, on the cultural context of emotion, and, finally, on the effects of individual speaker characteristics.

Section 2.4 explores affective speech synthesis, along with a final review of the current state of speech synthesis for Irish that covers some of the tools used in the present research. Finally, section 2.5 presents an overall summary of this literature review, and how the topics explored relate to the aims of the present research.

2.1. Theoretical Framework

2.1.1. Emotion

This subsection presents a brief review of the literature on emotion, starting with theories on where emotions originate from, its social functions, definitions of emotion for scientific purposes, and theories on how to categorize emotions for academic research.

2.1.1.1. *Theories of Emotion*

Expression of emotion is a system that guides social interactions and allows humans to “situate the self within a social and moral order” (Keltner, Sauter, Tracy, & Cowen, 2019, p. 133). There is a wide variety of theories that explore where emotions come from, those most often used in acoustic research coming from an evolutionary perspective based on the communicative function of emotion expression (Bryant, 2021).

This main approach used in emotion research is based on Darwin’s evolutionary theory, which considers emotion, reasoning ability, memory, and intelligence as adaptative behaviors from evolution in humans and other animals in the same way as anatomical structures. It views emotional expression in terms of its functionality, where expressive behaviors “act as signals and as preparations for action” and “(...) communicate information from one animal to another about what is likely to happen and thereby affect the chances of survival” (Plutchik, 1980, p. 5), considering the expression of emotion as a survival mechanism (Cornelius, 2000).

A similar view is expressed in the evolutionary psychology of Buss (1995), which aims to explain how humans and other creatures process specific kinds of information. In his words, it “conceives of the mind as containing a large number of specialized psychological mechanisms, each tailored to solving different adaptative problems, in addition to whatever somewhat more domain-general mechanisms it contains” (Buss, 2020, p. 317). Bryant (2021) also approaches vocal emotion from this adaptationist perspective, suggesting that “vocal signals are adaptations, (...) evolved solutions to recurrent adaptative problems of communication, shaped by natural selection”, and that “emotions can be construed as superordinate programs that coordinate physiological and psychological components to guide behavior in functionally organized ways” (p. 25). The expression of emotion can thus be considered a subcomponent of these programs, which communicate the speaker’s emotional state to other individuals.

Using evolutionary psychology to guide their review on the topic of human expression of emotion, Juslin and Laukka (2003) also mention that the key of the evolutionary approach rests in the concept that, in order to understand emotion, one has to understand its functions and “must consider how they reflect the environment in which they developed and to which they were adapted” (p. 771). This evolutionary approach with

a focus on communicative function has been increasingly applied to the study of vocal human emotion (e.g., Altenmüller, Schmidt, & Zimmermann, 2013; Bachorowski & Owren, 2003; Laukka & Elenberg, 2012; Pisanski, Cartei, McGettigan, Raine, & Reby, 2016; Sauter, Eisner, Ekman, & Scott, 2010; Shariff & Tracy, 2011), with an established research tradition tying together the acoustic structures of these adaptive behaviors and broad, discrete emotional categories (e.g., Banse & Scherer, 1996; Murray & Arnott, 1993; Russell, Bachorowski, & Fernandez-Dols, 2003).

In addition to Darwin's evolutionary view, Plutchik (1980) explores three other theoretical approaches in the study of emotion. One is the Psychophysiological tradition of William James and Karl Lange, which suggests that emotions come inherently from bodily changes that happen before the expression and are caused by a perception of an existing fact (James, 1884); i.e., what we call emotions is the experience of automatic evolutionary changes that our bodies have developed to features of the environment (Cornelius, 2000). Another approach is the Neurological tradition of Walter Cannon, (Cannon, 1929), who considers emotion-related physiological changes as bodily preparations for action (e.g., fight or flight), aligned to evolutionary theories of emotion. Finally, there is also the Dynamic tradition, based on the Freudian psychoanalysis idea that "some aspects of emotion are unconscious and cannot be examined by introspection alone" (Plutchik, 1980, p. 20). It contrasts the previous theories in the sense that there is no determined sequence of events, and that both the emotion and the bodily changes arise from an unconscious evaluation of one's situation.

Additionally, Cornelius (2000) presents two additional approaches to the study of emotion: the Cognitive Perspective and Social Constructivism. The Cognitive Perspective has as its central assumption the idea that emotion and thought are inseparable, i.e., emotions are highly dependent on an individual's *appraisal* (as defined by Arnold, 1960) of an event in the environment and its positive or negative implications for their personal wellbeing. On the other hand, Social Constructivism sees emotions as "cultural products that owe their meaning and coherence to learned social rules" (Cornelius, 2000, p. 5), focusing on the social role of emotion and culture-specific rules that govern how, when, by whom and, might we add, where/in which context and how strongly certain emotions should be experienced and/or expressed.

Further reviews on theories of emotion can also be found in Andersen and Guerrero (1998), Keltner, Oatley, and Jenkins (2019), Prinz (2004) and Scherer (1999, 2000b, 2000c). In the present research, emotion is viewed through the evolutionary lens – while the chosen method of data collection is not the most naturalistic (as detailed in section 2.2.2), the study focuses on a handful of emotions that, while considered relatively universally recognized across cultures (as will be explored in section 2.3), may also show some characteristics in their vocal expression that developed alongside of and are specific to the Irish language and dialect explored.

2.1.1.2. *Defining Emotion*

To conduct research on emotion, it is important to have a clear definition of the term and what it means as a concept before attempting to quantify the acoustics correlates of its vocal expression. This is, however, particularly challenging (Juslin & Scherer, 2005). Along with the many theories on the origins of emotion expression, emotions themselves are complex and dynamic patterns of behavior that involve multiple channels such as vocal cues, facial action and even bodily movement and posture (Keltner, Tracy, Sauter, Cordaro, & McNeil, 2016), which makes them hard to classify into discrete language labels due to the fluidity of the emotion expression process (Scherer, 2004). Patterns of emotion expression can also vary depending on the individual expressing the emotion, their history, social context, and relationship to the listener. This dynamism introduces further difficulty to the study of emotion, as Scherer (2003) mentions that, indeed, one of the main challenges in emotion research is precisely the lack of a consensual definition of the term.

For the purposes of the present research, a definition of emotion widely found in affective speech research outlined in Scherer (2003) and in Juslin and Scherer (2005) is used. They define *affect* as “an umbrella term that subsumes a variety of phenomena such as emotion, stress, mood, interpersonal stance, and affective personality traits” (p.69), during any of which a person is being affected by some event of behavior that causes their speech patterns to deviate from the normal (or neutral) state. These phenomena, or *affective states*, are then categorized through what the authors call the *design feature approach*, based on previous work by Scherer (2000c).

Juslin and Scherer (2005) suggest three broad classes of affective states: (1) emotions and stress; (2) moods and interpersonal stances; and (3) preferences, attitudes

and affect dispositions (e.g., personality). Emotion and stress are described as “(...) quite short but intense reactions to specific events of high pertinence to the individual” which “(...) produce reactions that entail a high degree of synchronization of all the subsystems (as detailed in Scherer, 2001), that is, coordinated changes in cognition, physiological arousal, motor expression, and motivational urge” (p.70). It is a spontaneous expression that can happen to an individual in isolation or around others. Examples of emotions are anger, sadness, joy, fear, elation etc.

In contrast, moods and interpersonal stances are rarely provoked by specific events or objects and not usually based on appraisal. Moods may occur for many different, sometimes unconscious reasons such as fatigue, hormonal balance and even weather (e.g., cheerful, or gloomy), while interpersonal stances relate to how an individual affectively interacts with someone else, which is usually long lasting or slow-changing, intentional or unintentional (e.g., cold and distant vs. warm and friendly), and depends on interaction.

Preferences and attitudes are defined as long-term affective evaluations of an object or a person (e.g., liking, desiring, valuing) and have low intensity and little impact on one's behavior as they are usually overridden by more situational factors (i.e., emotion or mood) which are stronger determinants of behavior (Scherer, 2000c). They are not quite the same as personality traits and affect dispositions, which are “characterized by very low affective intensity, very low synchronization, and behavioral impact usually felt only in interaction with situational variables” (Juslin & Scherer, 2005, p. 70). There is some research on the vocal markers of personality traits (e.g., Scherer & Scherer, 1981), but little is known about the vocal correlates of preferences and attitudes.

Another issue encountered in the study of vocal affect expression relates to the labelling of emotion categories. For that purpose, a review on the most prominent theories of emotion is presented, along with classification approaches used in the field.

2.1.1.3. Models for Emotion Research

Another issue that must be addressed in the research of affect and emotion is the labelling of emotions and affective states. Ideally, the meaning of words such as anger, happiness and sadness should be as clear and non-ambiguous as possible for scientific research, but the complexity of emotion expression also means that, despite a general

abstract consensus among humans on what those terms imply, there is a variety of ways in which these emotions can be experienced, expressed, and perceived.

In emotional speech research, most studies have utilized one of the following methods to categorize and label emotions: (1) the discrete approach; (2) the dimensional approach; and (3) appraisal theories.

The Discrete Approach (Basic Emotions)

The Discrete Approach to emotions, also called Theory of Basic Emotions, is based on Darwin's (1872/1998) evolutionary theories and developed by Tomkins (1995), Ekman (1992b) and Izard (1990, 1994). This approach is based on the theoretical existence of discrete emotions, which assumes that there are "(...) emotion-specific patterning for vocal as well as facial and physiological responding (...)" (Juslin & Scherer, 2005). It argues for the existence of a limited number of basic categories of emotion developed through the course of human evolution, suggesting that these "basic emotions" and the way they are expressed is the result of their adaptative value in relation to certain fundamental life tasks (Ekman, 1999) or to the functions they serve (Izard, 1993). While mostly focused on facial rather than vocal expression of emotion, their work (Ekman, 1982, 1992a, 1999; Izard, 1990, 1994) has shown the existence of cross-cultural, universally prototypical patterns of facial expressions for a number of basic emotion categories.

Ekman and Cordaro (2011) define basic emotions by suggesting that emotions can be called 'basic' if they present two characteristics: first, that they are discrete, i.e., they can be fundamentally distinguished from one another; and second, that these emotions "have evolved through adaptation to our surroundings" (p. 364). The number of basic emotion categories varies throughout the literature. Scherer (2000c) provides a range of categories between seven to fourteen, but also widely used in emotional research are the 'big six' emotions established by Ekman (1992b), which included anger, fear, happiness, sadness, disgust and surprise. These were later added upon (Ekman, 1999) to further include amusement, contempt, contentment, embarrassment, excitement, guilt, pride in achievement, relief, satisfaction, sensory pleasure and shame. The basic emotion approach has been widely used by studies on vocal expression of emotion, such as Juslin and Laukka (2003), and in speech technology research (Cowie et al., 2001).

Scherer (2000c) and Scherer, Johnstone, and Klasmeyer (2003) also support the idea of discrete categories of emotion with corresponding acoustic patterns due to the existence of semantic primitives of emotion (Johnson-Laird & Oatley, 1989), or words that denote emotional states. Scherer (2000c) mentions that such basic categories are a result of adaptive emotional strategies characterized by “specific eliciting conditions and its own physiological, expressive and behavioral patterns” (Scherer, 2000c, p. 147).

Despite often used in affective speech research, this approach has been criticized by other academics who consider the existence of basic emotions from which all other emotions can be derived as “an article of faith rather than an empirically or theoretically defensible basis for the conduct of emotion research” (Ortony & Turner, 1990, p. 315) that may not lead to progress on the field. Cowie and Cornelius (2003) point out that this small number of emotion families may contain emotions significantly different from each other, which leads to the issue of terminology for both researchers and listener judges. Emotions are often expressed, perceived and labelled differently depending on an individual’s culture, social norms and experiences, which may result in words used to describe certain emotions having different meanings across individuals and the voice qualities associated with emotions varying across cultures (Yanushevskaya et al., 2018).

For that reason, most recent studies focus on the dimensional approach to emotion, but there has also been some further, more recent exploration on the discrete approach. Recent interest in refining emotional categories is seen, for example, in Cowen, Elfenbein, Laukka, and Keltner (2019), also in evolutionary terms on a study of emotion recognition. As Bryant (2021) writes, because basic emotions loosely correspond to solutions to specific adaptative problems, the most ‘basic’ of emotions may be those more universally recognized.

The Dimensional Approach

Dimensional models categorize emotions as points within dimensional spaces. They can be unidimensional, e.g. based on activation (Scherer, 2000c), bidimensional, based, for example, on activation (passive-aroused) and valence (positive-negative) (Russell, 1980; Russell & Barrett, 1999), or multidimensional, based on, e.g., valence, level of activation and intensity (Plutchik, 1980; Russell, 1991; Schlosberg, 1954). While degree of arousal and

activation can both be defined as a deviation from one's normal psychological state (Russell, 1980), in many models, activation can also relate to what some may consider to be the intensity of the emotion, and to the physiological effects of arousal, along with valence. However, intensity can also be a dimension separate from activation as proposed by Daly, Lancee, and Polivy (1983), where it still relates to a deviation from a normal state, but in a non-directional manner (e.g., depression has negative valence and low activation, but high intensity).

Russell (1980) presents a two-dimensional circumplex model of affect that place emotions in a space determined by the dimensions of valence (misery-pleasure) and activation (arousal-sleepiness), with two additional opposite pairs presented around the circle: depression-excitement, and distress-contentment. In a later study, Russell (1991) also identifies three bipolar dimensions in categorizations of emotion across cultures: (a) evaluation, valence, pleasure, or positivity, (b) activity, arousal or activation, and (c) potency, power, control, or dominance. Despite the relative simplicity of the model, Russell (2003) argues for the importance of valence and arousal as the 'core processes' of affect, with this model having been extensively used in the study of emotion in speech (e.g., Bachorowski, 1999; Bachorowski & Owren, 1995; Bachorowski & Owren, 2003; Schröder, 2004a, 2004b; Schröder, Cowie, Douglas-Cowie, Westerdijk, & Gielen, 2001).

Another widely used dimensional model is Plutchik's (1980) multidimensional model, presented in figure 1 and based on his psychoevolutionary theory of basic emotions. It defines four dimensions (aggressiveness-awe, optimism-disapproval, love-remorse and contempt-submission) which determine eight primary emotions organized in a circle (ecstasy and grief, admiration and loathing, rage and terror, and amazement and vigilance), conceptualized as polar opposites that can be mixed within the dimensional space in order to create all other emotions. They can also present varying degrees of intensity (or level of arousal), shown by the vertical axis on the cone presentation, along with degrees of similarity, based on their proximity on the color wheel. It is therefore a model that classifies emotions both in terms of basic categories as well as variations of such categories within a dimensional space.

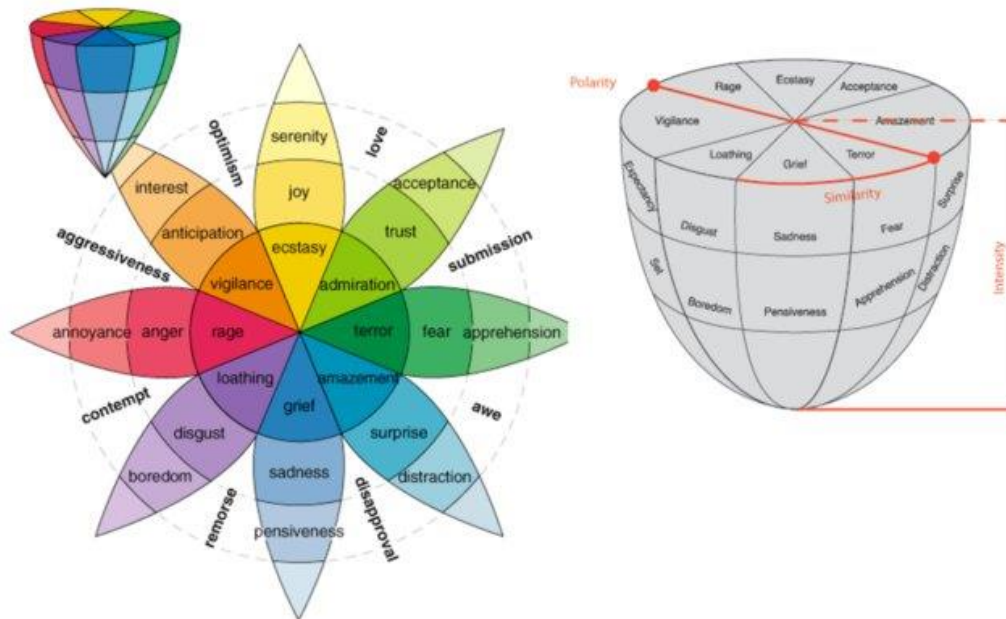


Figure 1 - Plutchik's Multidimensional Model of Emotion.
Retrieved from Vaughan (2011).

The idea of dimensional models can be considered to relate to the constructivist views presented previously. Additionally, L. F. Barrett (1998) also focuses on the dimensions of valence and arousal, but proposes that they are “important processes in determining whether a dimensional or a discrete emotion model best captures how people label their affective states” (p.579), as individuals with a higher focus on valence may fit a dimensional model of emotion better, while those with a higher focus on arousal fit a discrete model better.

Appraisal Theories

Appraisal theories, such as Scherer’s Component Process Theory (Johnstone & Scherer, 2000), argues that human expression is driven by cognitive appraisal of internal or external stimuli relevant to the organism (Scherer, 1986, 2004). This cognitive appraisal involves dynamic and multicomponent information processing that results in specific and context-sensitive ways of expressing emotion, both for the speaker in terms of vocal production (and any other physiological responses), as well as for the listener in terms of perception (Bryant, 2021). It considers emotion and the physiological changes caused by

emotion a result of the individual's evaluation of an event or object, and of how it affects their wellbeing.

Prototype Models

Prototype models, also known as lexical/meaning oriented models (Scherer, 2000c), suggest using the emotion lexicon of a language to discover the underlying structure of emotional phenomena. In other words, they are based on the idea that language and language-related structures shape how people perceive the world and process information (Guerrero, Andersen, & Trost, 1998). Fitness and Fletcher (1993) define prototypes as mental representations of categories of objects, events, and concepts like emotion. 'Prototypicality' can therefore be defined as "goodness of example" (Fehr & Russell, 1984), where emotion-denoting words can be ranked based on how accurately they exemplify an emotion archetype.

The prototype approach categorizes emotions in different families, each of them organized based on two dimensions. The vertical axis denotes the hierarchical relationship among three categories: the superordinate level, defined by valence; the basic level, which defines the core emotion within that category; and the subordinate level, which encompasses all the non-basic emotions related to the prototype. The horizontal axis determines the relationship between the members of each of the same category and considers both the contents of these individual categories as well as the hierarchical relations among them (Guerrero et al., 1998).

This approach has been criticized in terms of how emotion categories are created based on subjective accounts of emotion experiences (Clore & Ortony, 1991), and in terms of choosing which emotions qualify as a basic-level core emotion with its own separate profile (Fitness & Fletcher, 1993; Guerrero & Andersen, 1998). Like the discrete approach, boundaries between core emotion categories are also hard to define (e.g., Shaver, Schwartz, Kirson, & O'Connor, 1987), which leads many to suggest the need for revisions to the model.

The theoretical approach chosen for the present study is the discrete approach to emotions, which is closely tied to Darwin's evolutionary theories (Darwin, 1998) as the most 'basic' emotions are considered to more closely follow form-function relationships. The approach also has an established tradition in the field of research on the acoustic patterns

of emotion as adaptative behaviors (e.g., Airas & Alku, 2004; Banse & Scherer, 1996; Cowen, Elfenbein, et al., 2019; Cowie et al., 2001; Juslin & Laukka, 2003; Murray & Arnott, 1993; Russell et al., 2003; Sawamura et al., 2007; Scherer et al., 2003). While the validity and advantages of dimensional models are acknowledged for a broader investigation into the acoustic correlates of emotion expression, the choice for the discrete approach is also made based on practical considerations of the present research – a small number of discrete, well-separated categories of emotion was desirable for the creation of a simple set of changes applied to the voice source that could be tested among naïve listeners, and considered appropriate for the approach used for data collection (i.e., acted data, further explored in subsection 2.2.2). Additionally, perception tests are also conducted in Irish to avoid any issues with labelling of emotion categories, and translations were done with the assistance of a native Irish speaker.

2.1.2. Basics of Speech Production

This section explores the main theoretical concepts and frameworks on which the present research is based on, mainly the basics of speech production, voice source analysis and parametric models.

Vocal communication is traditionally divided into three levels: the production, the acoustic and the perception levels (Lieberman, Harris, Hoffman, & Griffith, 1957). Production refers to the physical setting of the vocal tract during speech production, which is divided into the phonation and the articulation phases. The phonation phase involves the expelling of air from the lungs through the trachea and the larynx by the action of the respiratory control muscles, making the vocal folds vibrate and produce sound. The articulation phase then describes the changes to the shape of the vocal tract, controlled by the muscles around the larynx that change the tension and position of the vocal folds. These folds control the airflow from the lungs, while the vocal tract filter boosts the frequencies that differentiate the sounds we produce at the acoustic level. Once speech sounds are expelled through the lips and nostrils, they then travel through the air and reach the listener's auditory system, which then processes the acoustic information at the perceptual level.

The Voice Source

In the production of speech, the delivery of air from the lungs is controlled by the muscles on the chest and abdomen. The air then passes through the trachea and the larynx, an adjustable, cartilaginous tubular organ (Reetz & Jongman, 2020) that controls the passage of air through the respiratory system like a valve (Kreiman & Sidtis, 2011). Also called the voice box, the larynx is where the vocal folds lie.

The vocal folds are two small, flexible protruding folds of tissue that run across the airway from below the thyroid notch to the arytenoid cartilages. The passage between the folds is called the glottis, which allows airflow from the lungs to pass through. As the vocal folds oscillate, they modulate the size of the glottis by alternately obstructing and opening it as they move back and forth. The periodic, oscillatory action of the folds converts the airflow from the lungs into a series of quasi-periodic pulses, which is what produces voiced speech. These pulses are what is called the glottal source, or the *voice source*. In contrast, unvoiced speech is produced when the airflow is constricted through the glottis, resulting in noise with no periodicity.

The rate of vibration of the vocal folds determines the fundamental frequency (f_0) of a sound (Johnson, 2011), which is perceptually correlated to pitch. This rate is also determined by the physical characteristics of the vocal folds, and varies depending on the speaker's gender, size, age and other individual characteristics.

The vocal fold vibration is rather complex, and described in detail in Kreiman and Sidtis (2011). Some known theories that try to account for the behavior of the vocal folds include Ishizaka and Flanagan (1972), Broad (1979) and Ishizaka and Matsudaira (1968). In general, the mode of vibration of the vocal folds is determined by the laryngeal tension settings of the vocal tract, controlled by the musculature and cartilages around the glottis including the vocal folds themselves (Laver, 1980). This interaction between the airflow and the tension settings changes the quality of the sound at the source and create what is known in the field as different *voice qualities*, a term used to describe perceptual auditory coloring of a speaker's voice in both individual and paralinguistic terms.

Voice Quality Frameworks

In a broad sense, *voice quality* refers to the long-term characteristics of an individual's voice and speech patterns. In a narrow phonetic use, the term describes the kind of voicing produced at the glottis, specifically at the vocal folds, to characterize different 'phonation types', or phonatory qualities (Esling, Moisik, Benner, & Crevier-Buchman, 2019, p. 2). In this sense, it can also describe short-term effects (i.e., down to the level of syllabic duration) produced by vocal tract configurations used in, for example, affect expression.

The description of voice quality uses auditory impressionistic labels which, in earlier research (Scherer, 1982, 1988), has been noted to produce long lists of terms that often conveyed different meanings to different researchers (Gobl & Ní Chasaide, 2003b; Juslin & Scherer, 2005; Laver, 1980; Ní Chasaide & Gobl, 2005). For this reason, a few frameworks have been developed for the study of voice quality that define a smaller number of labels based on the physical settings of the vocal tract required to produce certain phonation types.

Early work by Ladefoged (1971), referred to as the *continuum model of glottal states*, is an influential voice quality model of that focuses on describing the relationship between phonation types and glottal consonants. It explores how these linguistic categories are articulated exclusively at the glottis (not accounting for supraglottal articulation or aerodynamics), considered sufficient for the representation of glottal states between phonation types and consonants found across languages (Garellek, 2022, p. 7). His states of the glottis include *voice* (when vocal folds are vibrating), *voiceless* (when vocal folds are apart), *aspiration* (brief voiceless period after the release of a consonant, or a 'burst'), *murmur* ('breathy' voice), *laryngealization* ('creaky' voice), *glottal stop* (when vocal folds are held together) and *whisper* (when vocal folds are together or narrowed). The model is called a continuum because it considers these glottal states to fall into a continuum (with variations within each state), rather than being completely discrete categories.

This is further developed in a later update to the model (Gordon & Ladefoged, 2001), where these voice qualities are put into a continuum between when the folds are the most open (voiceless), followed by breathy, modal, and creaky voice, and up to when the folds are the most closed (glottal closure). The model is, however, mostly framed in terms of

glottal aperture (i.e., the opening between the arytenoids), and does not address muscular tension settings in detail. In this model, weaker forms of ‘breathy’ and ‘creaky’ voice are considered ‘lax’ and ‘tense’ respectively.

Another central framework in the study of voice quality is John Laver’s Voice Quality Framework (Laver, 1980). It consists of an in-depth description of laryngeal and supralaryngeal settings and how they relate to the production of different kinds of voice qualities, described in perceptual terms as voice quality categories that are still widely used in affective speech research.

The laryngeal settings, which are the ones more closely related to the present research and the main contributors to the perception of voice quality, are described in terms of the interaction between the pulmonic airflow and the tension of the muscles in and around the larynx. As air flows from the lungs through the glottis and makes the vocal folds open and close, these muscles control the degree to which the vocal folds tense or relax. The source signal, produced by the pulmonic airflow and modulated by the vibrating vocal folds, is then filtered by the supralaryngeal tract, and changes in its shape determine the resonant frequencies that produce different speech sounds.

Laver’s perceptual voice quality categories are *modal* (or neutral) *voice*, *whisper*, *creak*, *harsh voice*, *breathy voice*, and *falsetto*. *Modal voice* corresponds to a neutral state of the vocal apparatus that serves as a reference for all other categories, where “vibration of the true vocal folds is periodic, efficient, and without audible friction” (Laver, 1980, p. 94). *Whisper* is mainly characterized by turbulent air passing through the edges of this open cartilaginous glottis, producing the characteristic ‘whisper’ sound quality. *Creak*, or *vocal fry*, is characterized by a low fundamental frequency under that of the modal register and the f_0 and amplitude of the glottal pulses is very irregular, resulting in a tapping-like auditory quality. *Harshness* can only occur in combination with modal voice, resulting in harsh voice, also often called “pressed phonation” and presenting an aperiodicity in vocal fold vibration. *Breathy voice* involves constant glottal leakage and audible aspiration noise, characterized by limited vocal intensity and low f_0 . *Falsetto* is rarely explored in linguistic research, and beyond the scope of paralinguistic studies.

These categories are described in detail in Laver (1980) in terms of these muscular laryngeal settings, divided into three parameters: *adductive tension*, *medial compression*, and *longitudinal tension*. *Adductive tension* describes the tension of the interarytenoid muscles, which bring the arytenoid cartilages together and close both the cartilaginous and the ligamental glottis. *Medial compression* is defined as the compressional pressure on the vocal processes of the arytenoid cartilages, achieved through contraction of the lateral cricoarytenoid muscles and tension in the lateral parts of the thyroarytenoid muscles. Medial compression closes the ligamental glottis and may also close the cartilaginous glottis depending on the analytically separate adductive tension of the interarytenoid muscles. Lastly, *longitudinal compression* is achieved by the contraction of the vocalis and/or the cricothyroid muscles. A visual representation is seen in figure 2.

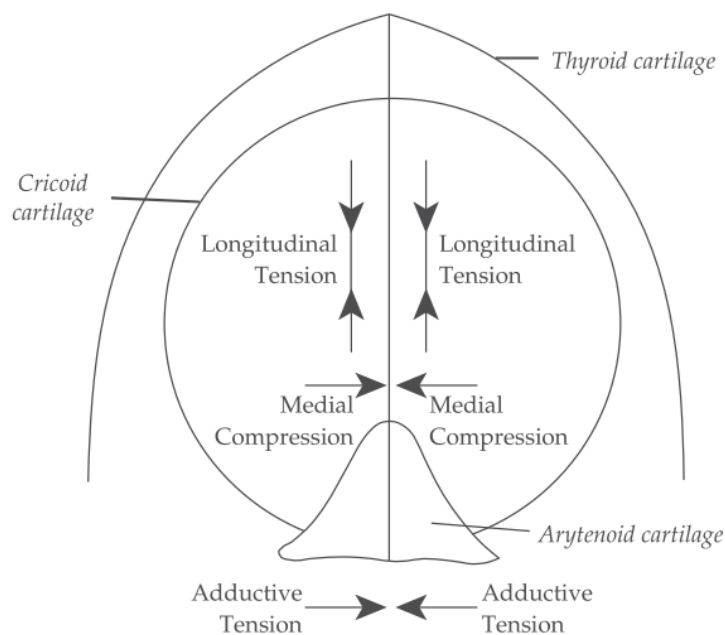


Figure 2 - The three laryngeal parameters of muscular tension. From Gobl and Ní Chasaide (2010), adapted from Laver (1980).

Additionally, many of these phonation types are not mutually exclusive and may occur in combination, forming compound phonatory qualities when the muscular settings of one category does not override the settings of another (i.e., when a potential

combination does not require two different tension settings for the same muscle). For example, one can combine whisper and modal voice or whisper and creaky voice, but breathiness and harshness are mutually exclusive as they require different types of vibration of the vocal folds (Gobl & Ní Chasaide, 2010; Laver, 1980).

In addition to these categories, Laver also mentions settings of overall degrees of muscular tension that affect the entire vocal system, namely *tense* and *lax* voice, which are of particular importance to the present study. These are believed to be mainly differentiated in terms of adductive tension and medial compression, along with subglottal pressure, and in terms of their acoustic differences (Laver, 1980). *Lax* voice is further described by Gobl and Ní Chasaide (2010) as a phonation “similar to breathy voice, sounding softer and lower-pitched than modal voice” (p.402).

Lastly, the *Laryngeal Articulator Model* (LAM) developed by Esling et al. (2019) is a more recent, broader model for the description of voice quality, which aims to provide a link between articulation and auditory qualities. It adopts Laver’s (1980) auditory voice quality categories but, rather than dividing the vocal tract into the laryngeal and supralaryngeal settings, the LAM instead splits it into the *upper (or oral) vocal tract* and the *lower (laryngeal) vocal tract*, the latter also called the “laryngeal constrictor” (see figure 3 for a schematic representation).

The LAM’s criticism of Laver’s model is that it is limited by using a “linear source-filter vocal tract model where the larynx is only a sound-source modulator and the tongue, jaw and lips (are) the principal consonantal and vocalic articulators” (Esling et al., 2019, p. 4), disregarding the role of the larynx as a complex structure of sources and resonators. Instead, it advocates for a model that links “glottal shape and oral articulation with the postural configurations (...) that occur between them” (and are independent) (id). In the LAM, the larynx is not just a source such as considered by the source-filter model (Fant, 1960/1971), but also acts as a filter through a complex articulator organ called the *laryngeal articulator*, located in the lower part of the vocal tract.

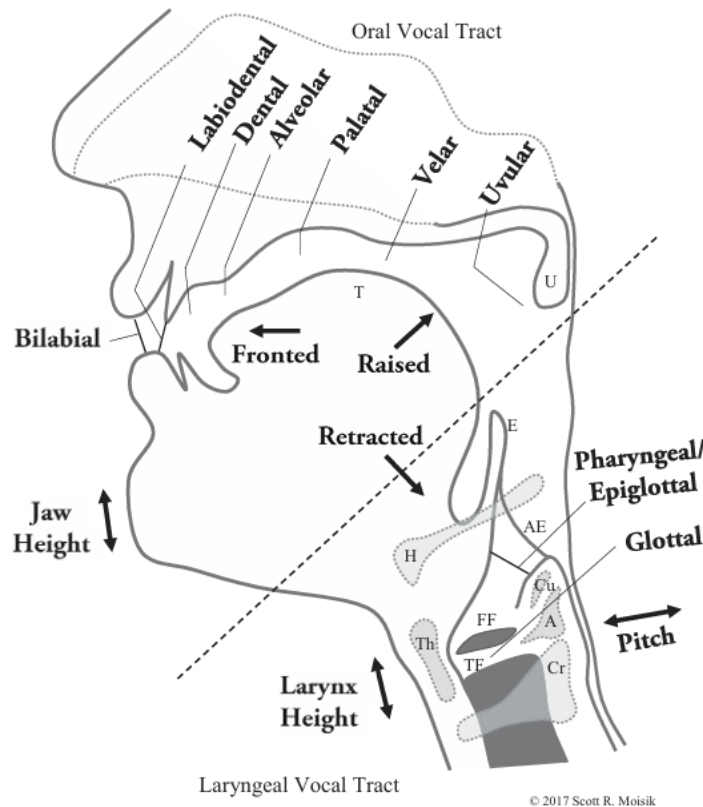


Figure 3 - Illustration of the Laryngeal Articulator Model, or the two-part vocal tract.

T = tongue; U = uvula; E = epiglottis; H = hyoid bone; AE = aryepiglottic folds; Cu = cuneiform cartilage; A = arytenoid cartilage; Th = thyroid cartilage; FF = ventricular (false) folds; TF = vocal (true) folds; Cr = cricoid cartilage.

Extracted from Esling et al. (2019), p. 6.

Laver's voice quality categories are therefore described differently in the LAM through an expanded set of "states of the larynx", expanding phonation above the glottis and beyond the intrinsic laryngeal musculature and revising the traditional notion of "states of the glottis" (Esling et al., 2019). *Modal voice* follows the description found in the previous literature (Hollien, 1974; Laver, 1980). *Breath* (and its voiced equivalent *breathy voice*, resulting from the addition of vocal fold vibration) is produced by a lack of engagement of the laryngeal constrictor, where the glottis is more open than in modal voice. *Whisper* (and *whispery voice*) is similar to breath at the level of the glottis, but with the addition of laryngeal constriction, causing more noise. *Creak* and *creaky voice*, or glottal fry, requires shortened and thickened vocal folds with a loose aspect (Esling et al., 2019) and differs further from Laver's description, where both *creak* and *creaky voice* are considered voiced and inherently have an f_0 , even if this f_0 is low enough for the glottal pulses to be perceived

independently. In the LAM, *creak* can be treated as voiceless because of its aperiodic pulses, while *creaky voice* can be treated as voiced because of its periodicity (Garellek, 2022). Harsh voice, like creaky voice, involves engagement of the laryngeal constrictor causing the voicing frequency to become irregular or aperiodic with variable amplitude, but doesn't share the low-pitched quality of creaky voice. Their laryngeal categories further include *breathing/inspiration*, *prephonation*, *glottal stop*, *epiglottal stop*, *falsestto*, *ventricular voice*, and *aryepiglottic trilling*.

Finally, Esling et al. (2019) also address the topic of tense and lax voices, mentioning that earlier research often regarded the term 'tension' as a generalized explanation for multiple sources of increased effort, tightness or force throughout the vocal tract, with no specified location. They therefore argue for a more detailed description to account for the many tension adjustments of the vocal tract through the laryngeal articulator of the lower vocal tract, taking the 'whole larynx' (as defined in Moisik & Esling, 2011), including the oral and laryngeal vocal tracts, into account in order to identify the sources of tension more accurately.

In the present research, for the purposes of creating a simple set of voice source transformations, vocal tract tension is condensed into the RD parameter, calculated through the transformed Liljencrants-Fant model (Fant, 1995; Fant, Liljencrants, & Lin, 1985) detailed in section 2.1.5.

The Vocal Tract

If we were to listen to the glottal signal itself, we would hear a something akin to a harmonically rich buzz rather than what we know as the human voice. This glottal flow needs to be shaped by the filtering action of the vocal tract for it to sound like the speech we hear in human communication (Kreiman & Sidtis, 2011).

The supraglottal vocal tract encompasses all the resonating cavities above the larynx (Raphael, Borden, & Harris, 2011), including the laryngeal cavity (larynx, or *voice box*), the pharynx, the oral cavity (mouth) and the nasal cavity (including the sinuses) (Stevens, 1999). This structure can be viewed as a filter, or an object with which the voice source interacts to produce the final sounds of a voice by a process called *resonance*. As with any object, the vocal tract has resonant frequencies at which they vibrate best depending on their size and

shape, which can change depending on the configuration of its articulatory organs (jaw, tongue, lips, the velum) and the settings of the larynx (which can be raised or lowered, changing the length of the vocal tract). When the frequency components of the voice source match the natural resonances of the vocal tract, the amplitude of these components are boosted (Kreiman & Sidtis, 2011).

Like the vocal folds, individual physical characteristics of a speaker also affects the properties of the vocal tract filter – men have on average longer vocal tracts, which also differ in terms of proportions of the oral and pharyngeal cavities compared to women and children (Fitch & Giedd, 1999). Accordingly, age also plays a part in shaping one's vocal tract. This, along with numerous individual characteristics, is what gives each person their own unique and recognizable voice.

2.1.3. Acoustic Theory of Speech Production

Because speech production is based on a series of interconnected and continuous physiological processes, researchers have found many approaches to simplify, measure and quantify the acoustic characteristics of speech.

In the most basic terms, speech is composed of a variety of complex soundwaves that can be decomposed into simpler sinusoidal components through the Fourier transform (Fourier, 1878). The frequency of the lowest component in this complex waveform (for voiced speech sounds) is the fundamental frequency (f_0), one of the main parameters used in the study of affective speech (Banse & Scherer, 1996; Johnstone & Scherer, 2000; Juslin & Scherer, 2005; Scherer et al., 2003). This frequency is the main determinant of the perceived pitch and is given by the rate at which the complex wave pattern repeats which, as mentioned, corresponds to the vibrating frequency of the vocal folds. In the Fourier transform, the positive integer multiples of the f_0 are called *harmonics*, which constitute the strongest frequency components in the signal after the fundamental. Depending on the shape of the vocal tract, other frequency components in the complex waveform (i.e., those matching the resonant frequencies of the vocal tract at a certain configuration) are also boosted in intensity. These frequency components boosted by the vocal tract configuration are known as the *formants* of a sound (Juslin & Scherer, 2005). Periodic waves produce voiced sounds such as vowels and voiced consonants, while aperiodic waves produce voiceless sounds such as stops and fricatives.

While acoustic information of the speech signal can be measured from speech recordings, the same cannot be said for the acoustic characteristics of the voice source. Because the physical components of the voice source are not visible and any internal measuring tools may affect voice production, their patterns are most often inferred from the speech signal (Scherer, 1986) by extracting the voice source parameters from the speech waveform. This conversion is done through mathematical models of speech production such as Gunnar Fant's (1960) *source-filter model* of speech production.

The source-filter model is one of the core models used in the speech sciences. It is based on the concept that "the speech wave may be uniquely specified in terms of source and filter characteristics" (p.15). In the frequency domain, it can be defined as follows:

$$S(f) = G(f)V(f)R(f)$$

Where $S(f)$ is the spectrum of the speech signal output (i.e., the frequency content and its respective amplitudes), $G(f)$ is the spectrum of the glottal source signal (i.e., $g(t)$ in the time domain), $V(f)$ is the transfer function of the vocal tract filter, and $R(f)$ is the radiation characteristics at the lips.

The glottal source signal provides a harmonic spectrum of frequency components, which act as the excitation signal. This type of source is involved in the production of vowels and voiced consonants. Other sources involve aperiodic noise resulting from a turbulent airflow at some point of constriction in the vocal tract.

The filter $V(f)$ then imposes its resonant characteristics onto the source signal to produce the sounds we hear as speech (Styger & Keller, 1994). As mentioned, the vocal tract configuration boosts certain resonant frequencies (the formants) of the glottal flow spectrum. Most vocal tract configurations can be modelled through complex conjugate pairs of poles and zeroes, where the poles in the model represent these natural frequencies of the vocal tract (Stevens, 1999). All-pole models are widely used to estimate formants in vowels (e.g., Alku, Pohjalainen, Vainio, Laukkanen, & Story, 2013). Formant resonances are defined by both their central frequency and their bandwidth, which is the damping rate of the formant resonance. The number of formant resonances is, in theory, infinite, but only

the first few (those with the lowest resonant frequencies) are considered perceptually relevant (Kreiman & Sidtis, 2011). The combination of these formants constitutes the transfer function of the vocal tract.

The final component of the model is the *radiation characteristic* $R(f)$, which describes the process of the sound dispersing in all directions through the air after leaving the vocal tract system through the lips. The effect of the sound radiation at the lips on the speech output spectrum is that the amplitude levels of higher frequency components increase relative to those of lower frequency components at a rate of approximately 6 dB per octave.

This filtering process of the source signal results in a speech output signal which contains salient acoustic parameters. A listener's ability to perceive differences in the speech signal resulting from changes in these parameters is how they are able to not only differentiate between sounds, but also to identify emotional information (Bryant, 2021; Kreiman & Sidtis, 2011).

This model, however, still has its limitations. For instance, it does not take any subglottal interactions into account, considering the glottis as the start of the process. Furthermore, it assumes that the source and the filter are independent and non-interactive components. Nevertheless, this model is widely used in research and practical speech applications.

2.1.4. Voice Source Analysis

As mentioned, measuring voice quality and voice source data in general is a difficult task that requires highly specialized tools and presents many challenges given their internal, hard to access nature. Several techniques have been developed for analyzing and extracting measurable parameters from the voice source, from direct measures to mathematical modelling. These techniques are explored in this section.

One method for directly estimating the glottal source was developed by Sondhi (1975), in which a speaker produces a neutral vowel into a reflectionless tube consisting of a fiberglass cone inside a brass pipe. This cone dampens the resonances of the vocal tract, removing the need to use inverse filtering techniques (i.e., methods of estimating and removing the effects of the vocal tract transfer function to obtain an estimate of the glottal

source, further explored below), and the pressure waveform is captured by a microphone placed within the tube and seen as an approximation of the glottal volume velocity. While it is a non-invasive technique that also does not require a sound-proof environment due to the location of the microphone, it is also limited to non-nasal vowels and not applicable to pre-recorded audio. Another way to obtain the glottal volume velocity is through a circumferentially vented pneumotachograph mask (Rothenberg, 1973), used for acquiring the oral flow at the lips by detecting pressure drop across a resistance (e.g., a wire screen) (Rothenberg, 1972). This technique, however, still requires inverse filtering to obtain the voice source signal (i.e., the glottal volume velocity).

Another technique is that of semiconductor miniature pressure transducers, inserted between the vocal folds through a thin catheter to measure subglottal pressure (Cranen & Boves, 1985; Kitzing & Löfqvist, 1975). This method is, however, very invasive and uncomfortable for the subject. Adding an object to the vocal tract also interferes with vocal production, thus not being appropriate for naturalistic voice studies.

Indirect measures of the voice source (i.e., related to the spectral characteristics of the speech signal) have also been explored, such as electroglottography (Johnstone & Scherer, 1999; Laukkanen, Vilkman, Alku, & Oksanen, 1996). They however tend to be hard to interpret depending on the degree of contact of the vocal folds. Measures from the speech output spectrum such as spectral energy distribution (i.e., how the energy of the spectrum is spread across different frequencies, as done by Guzman, Correa, Muñoz, & Mayerhoff, 2013) and measures related to pulse-to-pulse stability such as jitter and shimmer (random variations in f_0 and fluctuations in pulse-to-pulse amplitude, respectively) have also been used in the study of the relationship between emotion and the voice source (Johnstone & Scherer, 1999). For example, high levels of jitter and shimmer have been suggested to correlate with hoarse voice but are not the only parameters that might present instabilities (Gobl & Ní Chasaide, 2010). Additionally, one of the main parameters reliably reported to differentiate phonation types is spectral tilt, i.e., the overall slope of the frequency spectrum (id.). However, because these methods do not separate the effects of the source and the filter very well, the source measures acquired are not particularly accurate (Ní Chasaide & Gobl, 1997).

In addition, Mehta, Espinoza, Van Stan, Zañartu, and Hillman (2019) used miniature high-bandwidth neck-surface accelerometers to acquire measures of laryngeal voice production. While the approach was previously used to extract fundamental frequency in noisy environments (Porter, 1963) and expanded to temporal, spectral and cepstral measures (e.g., Castellana, Carullo, Corbellini, & Astolfi, 2018; Titze, Svec, & Popolo, 2003), their study examines a “low-bandwidth measure of spectral tilt that can be directly estimated from the accelerometer signal” (p. 387), i.e., the difference between the spectral amplitude (in decibels) of the first and second harmonics (L_1-L_2). In acoustic analysis, this measure has been commonly related to pulse skewness of glottal airflow, open quotient (i.e., the proportion of the glottal cycle during which the glottis is open), and medial vocal fold thickness (Heinrich, d'Alessandro, & Doval, 2001; Klatt & Klatt, 1990; Swerts & Veldhuis, 2001; Zhaoyan Zhang, 2016).

Their results showed a high degree of correlation between measures extracted from the accelerometer signals and glottal flow estimates. Mehta et al. (2019) argue that this method supplants acoustic voice measures as it is less sensitive to environmental noise (in contrast to a microphone) and, because it is placed on the speaker’s anterior neck surface and below the larynx, it is also not sensitive to articulatory modulations of the acoustic speech signal. This, along with the low-profile appearance of the device, also makes it more appropriate for collection of naturalistic voice recordings (i.e., acquired outside of a laboratory setting) during a speaker’s day-to-day activities. This approach was also used by Włodarczak, Ludusan, Sundberg, and Heldner (2025) aiming to provide a better estimate of the voice source than would be achievable from the acoustic analysis of raw audio, but no advantages for one method over the other were found in discriminating between the phonation types tested.

Most studies follow the principles of the source-filter theory by using mathematical models to estimate voice source parameters from the acoustic speech waveform. This usually involves high quality, phase linear audio recordings, along with inverse filtering techniques.

Inverse filtering

Inverse filtering is a voice analysis technique that removes the effects of the filtering action of the vocal tract by separating the source and the filter. It does so by running the signal through a transfer function that cancels the effect of the vocal tract and provides an estimate of the voice source (Fant, 1995; Gobl & Ní Chasaide, 2003b). While the definition seems straightforward, inverse filtering is still inherently difficult to conduct, as this separation of source and filter from the speech signal can only be done based on current knowledge of the anatomy and functionality of the voice source and the filter, as the only objectively known measure is the recorded speech signal itself.

To analyze large amounts of data, automatic and semi-automatic inverse filtering methods have been developed, most based on linear predictive analysis (e.g., Alku & Vilkman, 1994; Fröhlich, Michaelis, & Strube, 2001; Wong, Markel, & Gray, 1979). The most widely used algorithm is the Iterative Adaptive Inverse Filtering (IAIF) (Alku, 1992) such as, for example, in Airas, Pulakka, Bäckström, and Alku (2005). However, even though some studies based on such techniques have provided some information on the voice source, they tend to perform much less effectively on voice qualities different from modal voice, i.e., when there is no true closed phase of the glottal cycle (Gobl & Ní Chasaide, 2003b). Through the past few years, these methods have been actively modified and improved upon (Mokhtari, Story, Alku, & Ando, 2018; Perrotin & McLoughlin, 2017; Ziqian Zhang, Xiao, & Huang, 2022). Despite such improvements, however, automatic systems do not yet perform at a high enough level of accuracy necessary for highly detailed analysis of glottal waves, and require high quality recording equipment in ideal, sound-proofed environments.

Alternatively, there are also interactive manual inverse filtering methods which allow for a much more accurate, refined estimation of the voice source signal parameters (e.g., Carlson, Granström, & Karlsson, 1991; Gobl & Ní Chasaide, 2010; Ní Chasaide & Gobl, 1997; Ní Chasaide, Gobl, & Monahan, 1992). Such manual methods, however, require a high level of expertise and are very time-consuming, being therefore more suitable for micro-studies and the analysis of smaller amounts of data. The manual analysis software used for the present research (Gobl & Ní Chasaide, 2010; Ní Chasaide et al., 1992), for example, requires each pulse of an utterance to be individually analyzed, first by cancelling formants and then by matching the pulse of a model to the acquired differentiated glottal flow and

the desired voice source parameters. The specific parameters vary depending on the model used for this matching process.

Model Matching

There are many available approaches for parametrization of the glottal flow waveform. While some techniques simultaneously estimate filter and source parameters (Fröhlich et al., 2001; Kasuya, Maekawa, & Kiritani, 1999), the method most often used to parameterize the voice source is *model matching*, i.e., adjusting a parametric source model to the shape of the glottal pulses acquired from the inverse filtering process to capture its characteristics based on a set of parameters (Fant et al., 1985; Fujisaki & Ljungqvist, 1986; Klatt & Klatt, 1990; Rosenberg, 1971; Veldhuis, 1998). Automatic model matching methods also exist, optimized for both the time domain (e.g., Strik & Boves, 1994) and the frequency domain (e.g., Olivera, 1993). Problems with estimating the voice source during inverse filtering, however, carry over to model matching techniques, with manual techniques often being more robust than automatic methods, but also much slower and with higher expertise requirements. Details on such models used for voice source parametrization are explored below.

2.1.5. Glottal Source Models

This section presents a brief overview of the most well-known glottal source models. A thorough review can be found in Cummings and Clements (1995). Additionally, Kreiman, Garellek, Chen, Alwan, and Gerratt (2015) also introduce the models detailed here in the form of a perceptual evaluation of their performance for synthesized stimuli.

The Liljencrants-Fant Model

Of particular importance to this research is the *Liljencrants-Fant Model* (Fant et al., 1985), which consists of a glottal flow model with four independent parameters. It models the differentiated glottal flow (i.e., the glottal flow derivative) in the time domain.

The model determines the shape of the glottal pulse (i.e., the result of a single opening and closing cycle of the vocal folds) in terms of four parameters. Three of them pertain to the frequency, amplitude, the exponential growth constant of a sinusoid (E_0 , α and ω_g , detailed below), the fourth being the time constant of exponential recovery (i.e.,

the return phase TA , or the time from the maximum closing discontinuity to the maximum closure of the vocal folds) (Fant et al., 1985). Along with f_0 , the parameters in the model tend to be mutually dependent shape parameters related to the voice source.

The model consists of two segments. The first segment includes the first three parameters mentioned and represents the open phase of the glottal cycle, modelling the differentiated glottal flow from the glottal opening, t_0 , to the point of maximum excitation, t_e . It is represented by an exponentially growing sinusoid defined by three of the parameters, as seen in figure 4 below.

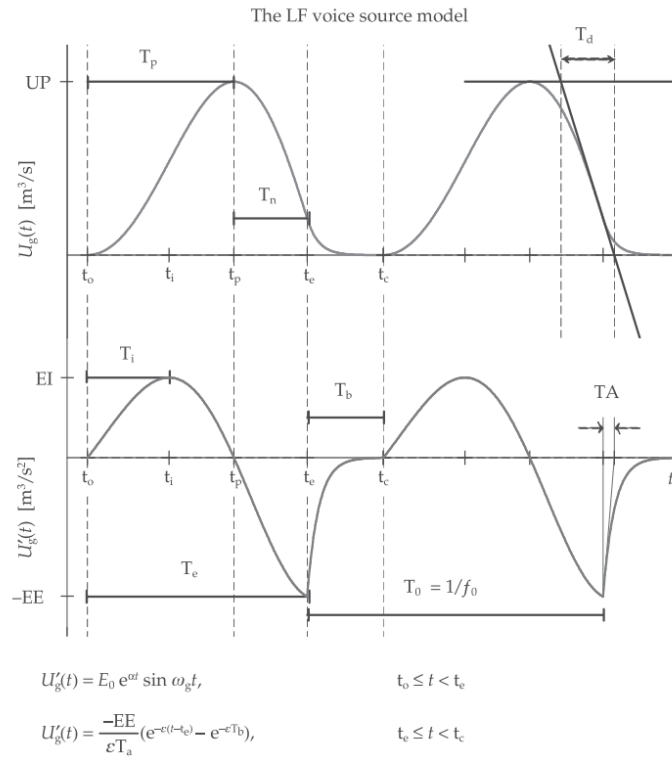


Figure 4 - The LF voice source model. The top graph shows the glottal pulses; the bottom graph shows the corresponding LF model pulses. Extracted from Gobl and Ní Chasaide (2010).

The E_0 parameter is generally determined by the EE parameter (see below) and can be determined by the formula: $E_0 = -E_e e^{-\alpha T_e} \csc \omega_g T_e$, where $\alpha = -B\pi$, and where α determines the rate of amplitude increase with B representing the negative bandwidth of

the exponentially growing sinusoid. α is typically determined by the other parameters in order to achieve ‘area balance’ (see further below), and $\omega_g = 2\pi F_g$, where $F_g = (1/2t_p)$, and where F_g is the frequency of the sine function and t_p is the time from glottal opening to maximum airflow (Gobl & Ní Chasaide, 2010).

The second segment corresponds to the return phase, which models the differentiated glottal flow from the time point of the main excitation, t_e , to the time of glottal closure t_c . It consists of an exponential function that represents the residual airflow after the main excitation, when the vocal folds close, and determines *dynamic leakage*, or the residual airflow through the glottis when the vocal folds close. The parameter of this segment is the time constant TA , which is obtained by the projection of the time axis on the tangent at time point t_e and measures the “effective duration” of the return phase (Gobl & Ní Chasaide, 2010).

As mentioned, these four parameters are the basis by which the LF model pulse shape is determined, and the model itself is further characterized by a requirement of area balance, i.e., the positive area of the LF pulse should be equal to its negative area (Ní Chasaide & Gobl, 1997). However, the following parameters described below are often used with the LF model to characterize the glottal pulse.

In the LF model, f_0 is defined as $1/T_0$, where T_0 is the duration of the glottal cycle measured between the main excitation of two consecutive glottal pulses.

The EE parameter, or the *excitation strength*, determines the overall intensity of the signal. It is described by Gobl and Ní Chasaide (2010) as the negative amplitude of the differentiated glottal flow at the main excitation, which occurs at the instant of glottal closure. They state that “(...) at the production level it is determined by the speed of closure of the vocal folds and by the airflow through them” (p.389), and that it corresponds closely to the overall intensity of the source signal at the acoustic level. In practice, the higher the EE, the higher the amplitude of the harmonics, and thus the higher the perceived loudness.

RA corresponds to the *dynamic leakage* of the return phase, defined as TA/T_0 and effectively measuring the duration of the return phase. It relates to the sharpness of the glottal closure, being low when vocal folds close more instantaneously and high when they close more gradually with more air leakage, and to spectral tilt (the degree to which the

amplitude of spectral components decreases as the frequency increases). A high RA tends to indicate breathy voice, while a low RA happens in tense voice.

RK is a measure of glottal pulse symmetry, or *glottal skew* – the relationship between the times of the rising and falling branches of the glottal pulse. The higher the RK, the more symmetrical the pulse. A more symmetrical pulse shape boosts the lower harmonics and also deepens the spectral dips of weakened or missing harmonics (Gobl & Ní Chasaide, 2010). Lower RK values indicate a more skewed glottal pulse, and typically a tenser phonation.

RG corresponds to the glottal frequency F_g normalized to f_0 ($RG = F_g / f_0$), where F_g is the characteristic frequency of the glottal pulse during the open phase. RG captures the frequency of the boosting effect of the lowest frequency components of the glottal pulse and tends to vary inversely with OQ and UP. A high RG corresponds to attenuated levels of the lowest end of the source spectrum (Gobl & Ní Chasaide, 2010).

UP is the peak amplitude of the glottal flow pulse. High UP values correspond to increased amplitude of the lowest harmonics in the source spectrum, which is often seen in breathy voice.

OQ is the *open quotient*, mentioned previously as the proportion of the glottal cycle for which the glottis is open. It can be defined through other parameters, as $OQ = T_e / T_0$ or $OQ = (1 + RK) / (2RG)$, and mainly affects the amplitudes of the lower components in terms of the source spectrum.

Finally, RD, proposed by Fant (1995, 1997), is a global waveshape parameter considered “the most effective single measure for describing voice qualities” (Fant, 1995, p.152). It is closely related to the effective pulse declination time $T_d = UP / EE$ and mathematically defined as:

$$RD = 1000 \times \left(\frac{UP}{EE} \right) \times \left(\frac{f_0}{110} \right) = \frac{1}{0.11} \times \left(\frac{T_d}{T_0} \right)$$

Therefore, RD is T_d normalized to the fundamental period, scaled by the factor of 0.11^{-1} so that its numerical value corresponds to T_d in milliseconds for an f_0 of 110 Hz (Fant,

1995). As mentioned previously, it relates to vocal fold tension (i.e., voice quality variation on the tense-lax continuum): a high RD value generally corresponds to a relaxed, lax voice quality, and a low RD value corresponds to a tenser, harsh voice quality.

There are also a few additional parameters not included in the original LF model, but which often appear in voice source research. The AH parameter, or *aspiration noise*, refers to the turbulent noise produced at or near the glottis in addition to the quasi-periodic sound source produced in phonation. Although the noise source is not often included in voice source models, Gobl and Ní Chasaide (2010) mention that a noise generator can be used with a model to provide the noise component of the voiced excitation. A version of the LF model which incorporates amplitude modulated aspiration noise is presented in Gobl (2006).

Along with RD, there are also two other global amplitude-based parameters often used in related research. One of them is the AQ parameter, an amplitude-domain quotient for parametrization of the glottal source that is essentially the same as $T_d = UP/EE$. It is described as the ratio between the amplitude of the AC-flow of the glottal waveform (i.e. the AC-amplitude of the flow signal, or the largest peak-to-peak amplitude, which is equivalent to UP) and the amplitude of the minimum of the differentiated flow (the level of the negative peak amplitude of the flow derivative, or the largest amplitude of the cycle-to-cycle minimum derivative, which is effectively the same as EE) (Alku & Vilkman, 1996; Campbell & Mokhtari, 2003).

Alku, Bäckström, and Vilkman (2002) have also proposed the NAQ parameter, or the normalized amplitude quotient. NAQ is the AQ parameter normalized to the length of the fundamental period, taking f_0 into account, and is similar to RD excluding the scaling factor: $NAQ = AQ/T_0 = 0.11 \cdot RD$ (Gobl & Ní Chasaide, 2003a), and correlates well with the tense-lax dimension of the voice source (Gobl & Ní Chasaide, 2010). It tends to be higher for breathier, laxer voice (i.e., a smoother glottal flow) and lower for pressed, tenser phonation (i.e., a shorter glottal closing phase and high vocal effort) (Alku et al., 2002). It has also been suggested that AQ better captures the tense-lax variation of the voice source across a larger f_0 range than NAQ, which is therefore more suitable for more limited f_0 ranges (Gobl & Ní Chasaide, 2003a).

A detailed discussion of the ways in which these parameters covary can be found in Gobl and Ní Chasaide (2010). A few examples are as follows. The return phase varies inversely with EE – the weaker the excitation strength, the higher the RA. Conversely, RA and RK also covary, and therefore a longer return phase and lower EE result in a more symmetrical pulse shape. RG varies inversely with OQ and UP, while OQ correlates with RK and RG. Gobl and Ní Chasaide (2010) also mention how these parameters can be used to characterize Laver’s phonation types. For example, breathy voice shows high dynamic leakage (high RA) with a more symmetrical glottal pulse (high RG and OQ). Whispery voice is like breathy voice in terms of source measurements, but more extreme, with the main distinction being a lower RK and OQ which results in a more skewed glottal pulse. Lax voice, on the other hand, is also similar to breathy voice in terms of phonation, but closer to modal voice. Finally, tense voice shows low RA (corresponding to a sharp, full closure of the vocal folds), low OQ and high RG.

One of the main criticisms to the LF model is its missing frequency domain modelling. Because its parameters and functions are defined in the time domain, its implementation is subject to aliasing distortion, as well as it not being quite as accurate a representation of the real human voice source waveform. To address this, a few studies have proposed frequency domain descriptions of the model (Aicha, 2018; Gobl, 2021), which have been shown to respectively reduce distortion (Z. Wang & Gobl, 2023a) and improve the quality of the synthesized speech (Aicha, 2018).

Rosenberg Model (Rosenberg, 1971)

The Rosenberg Model (Rosenberg, 1971) is a trigonometric model that describes the opening and closing phases of the glottal flow volume velocity through two timing parameters and one amplitude parameter. The time parameters T_p refers to the time from the onset of the glottal pulse to its peak (opening phase), while T_n refers to the time of the peak of the glottal pulse to its offset (closing phase), and the amplitude parameter α refers to the maximum amplitude value of the glottal pulse (Cummings & Clements, 1995). It is defined by the formula below:

$$u_g(t) = \frac{\alpha}{2} \left(1 - \cos \left(\frac{t\pi}{T_p} \right) \right) \quad \text{for } 0 \leq t \leq T_p$$

$$u_g(t) = \alpha \cos \left(\frac{\pi}{2} \frac{t-T_p}{T_n} \right) \quad \text{for } T_p \leq t \leq T_p + T_n$$

The glottal pulse $u_g(t)$ is assumed to be zero during the glottal closure. This model has been considered easy to utilize and to generate good synthetic speech (Holmes, 1973), and achieved the highest preference scores among a set of different pulse shapes in comparison to its corresponding polynomial model (Rosenberg, 1971).

Fujisaki-Ljungqvist Model

The Fujisaki-Ljungqvist model (Fujisaki & Ljungqvist, 1986) is a six-parameter model that defines the shape and duration of the flow derivative waveform through a set of four polynomial segments. Its parameters consist of three timing parameters, W (open phase duration), S (pulse skew) and D (time from glottal closure to the maximum negative flow); and three amplitude parameters, A (the amplitude slope at the moment of glottal opening), B (the amplitude slope immediately before the glottal closure) and C (amplitude slope following the glottal closure). These are illustrated in figure 5.

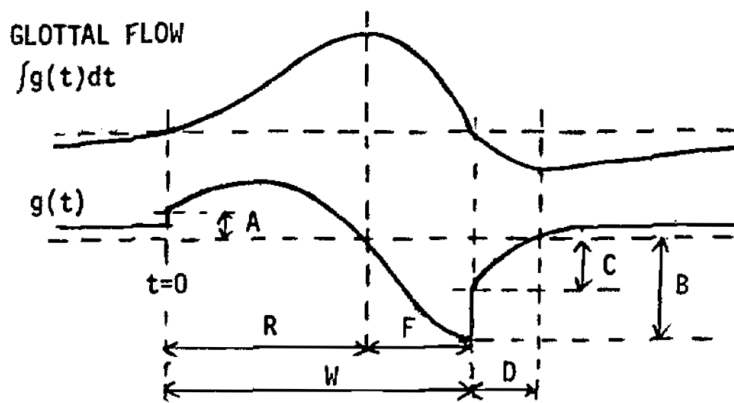


Figure 5 – The glottal pulse in the Fujisaki-Ljungqvist model. From Fujisaki and Ljungqvist (1986).

W and S are described in terms of R and F , which correspond to the open pulse duration segments before and after the glottal peak, respectively:

$$W = R + F, \quad S = \frac{R + F}{R - F}$$

The fundamental period T is specified separately, based on the model's individual polynomial segments. The model defines the glottal flow derivative $u'_g(t)$ based on the value of t relative to the R , W , D and T parameters. This derivative also incorporates the effects of lip radiation, which causes a change in the spectral slope of 6dB per octave.

A unique feature of this model is the possibility of controlling the slope of the waveform at the time of the glottal opening through the amplitude parameter A , allowing for the inclusion of discontinuities at the glottal opening and closing. However, it is also considered more difficult to use than other models because of its more complex functionality (Cummings & Clements, 1995).

Other Glottal Source Models

Some alternative glottal source models based on those previously described are the Rosenberg++ model (Veldhuis, 1998), and the Shue-Alwan model (Shue & Alwan, 2010), later modified in Chen, Shue, Kreiman, and Alwan (2012).

The Rosenberg++ (R++) model (Veldhuis, 1998) is an extension derived from the original Rosenberg model. It uses the same set of time and R parameters of the LF model and is presented as a more computationally efficient alternative to it. It adds a return phase to the model to improve its flexibility, allowing the user to specify the T_p parameter which, in the original model, is fixed as $t_p = 2t_e/3$. It has been shown to produce synthetic speech that is perceptually equivalent to that generated by the LF model but is still less frequently used than the LF model at the time of writing.

Shue and Alwan (2010) proposed a time-domain glottal flow model (not the flow derivative) derived from the integrated LF model. It models the glottal area waveform based on observations made from high-speed endoscopic recordings of the laryngeal vibrations instead of using flow pulses or derivatives. It consists of five control parameters: T_0 (the fundamental period), OQ (the open quotient), α (asymmetry coefficient), S_{op} (speed of the

opening phase) and S_{cp} (speed of closing phase). Like the LF model, it uses a combination of sinusoidal and exponential functions but has a higher level of flexibility in adjusting the slopes of the opening and closing phases separately. The glottal flow is defined in terms of these control parameters and time intervals, as illustrated in figure 6:

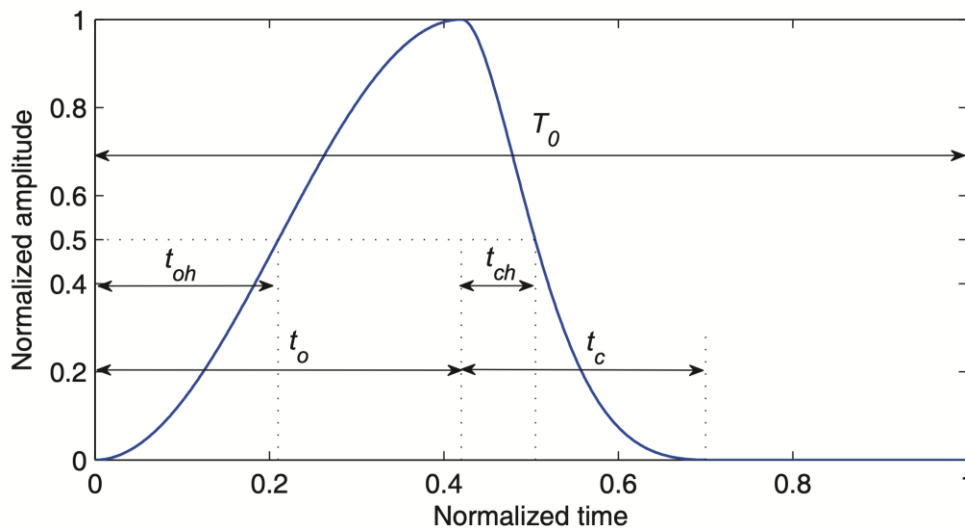


Figure 6 - Illustration of the Shue-Alwan model glottal pulse. From Shue and Alwan (2010).

With the control parameters defined as:

$$OQ = \frac{t_o + t_c}{T_0}, \alpha = \frac{t_o}{t_o + t_c}, S_{op} = \frac{t_{oh}}{t_o} \text{ and } S_{cp} = 1 - \frac{t_{ch}}{t_c}$$

Where T_{oh} and T_{ch} are at 50% of the amplitude.

Later, Chen et al. (2012) proposed modifications to the Shue-Alwan model by redefining the speed of the opening and closing phases, allowing for lower computational complexity, faster waveform generation, and more accurate pulse shape manipulation.

The Shue-Alwan model has been shown to outperform the LF model in terms of accurate estimation of the source signal (Shue & Alwan, 2010), and, along with the LF model, to outperform the Rosenberg and FL model in terms of the quality of the synthetic stimuli generated (Kreiman et al., 2015).

In the end, the LF model was chosen for the current research for multiple reasons. First, it is widely used in the literature, and many of the tools available for conducting the analysis and synthesis of speech hereby proposed were developed with the LF model's parametrization as a basis (e.g., Murphy, Yanushevskaya, Ní Chasaide, & Gobl, 2020; Ní Chasaide et al., 1992; Z. Wang & Gobl, 2023b). Results from Kreiman et al. (2015) regarding its performance in creating synthesized stimuli was also a motivator. Additionally, the global waveshape parameter RD is appropriate for our goal of creating a limited set of simple, systematic manipulations to a synthetic baseline instead of wider range of glottal source parameters.

2.1.6. Intonation and Prosodic Structures

This subsection presents an overview of prosody and intonation, as part of the aim of this study is to examine the relationship between shifts in voice source parameters and affect in relation to the prosodic structure of Irish utterances.

Arvaniti (2020) defines prosody as an “umbrella term used to cover a variety of interconnected and interacting phenomena, namely stress, rhythm, phrasing and intonation” (p.1), which relate to the dynamic, temporal patterning of the voice source. Although earlier studies on prosody were less concerned with individual phonetic segments of speech (i.e., segmentals, such as vowels or consonants) and more with the properties of suprasegmentals (i.e., prosodic features such as pitch, intensity, duration and voice quality) and other larger elements like syllables or larger units of sounds, Arvaniti (2020) states that “prosody does not operate independently of segments; rather, it profoundly affects segment realization” (id.), arguing that prosody is an essential element in any kind of phonetic research. While paralinguistic information can be conveyed without the linguistic signal, such as in languages unknown to the listener (Ladd, 2008) or even through short bursts (Patel, Scherer, Björkner, & Sundberg, 2011), prosody and paralinguistics are often conflated because the acoustic parameters used to encode linguistic prosody are the same as those used to convey paralinguistic information, e.g., fundamental frequency (Arvaniti, 2020).

The phonetic dimensions of prosody include: (1) f_0 , perceived as pitch; (2) temporal factors such as duration (of syllables or other suprasegmental parts of speech) and rhythm, related to the impression of ‘beat’ we hear in running speech; (3) intensity, or, perceptually,

loudness; and (4) voice quality, which involves the entire source signal. While voice quality is not as broadly explored in the field of prosody as the other three dimensions, research has shown that it not only plays a major role in the signaling of affect (Gobl, Bennett, & Ní Chasaide, 2002; Gobl & Ní Chasaide, 2003b), but also that the dynamic modulation of the voice source is suggested to contribute to prosodic signaling in general, i.e., intonation, as outlined below.

Jones (1960) defines intonation as “the variations which take place in the pitch of the voice” (p.135) in connected to speech, or, in other words, “language-specific and systematic modulations of fundamental frequency that span entire utterances and have grammatical function(s)” (Arvaniti, 2020, p. 19). These variations serve as a “means of conveying information in speech which is independent of words and their sounds” (Nolan, 2006, p. 433), which may be: (1) the grammatical structure of a sentence, or the ‘punctuation’ that marks the divisions between grammatical units like compound sentences or phrase boundaries; (2) conversation regulation and turn-taking between speakers; (3) information structure or focus, which is a means of emphasizing a particular piece of information that the speaker wishes to focus on, or relative to the information already shared in the conversation; (4) discourse function or sentence type, i.e., the differentiation between statements, wh-questions, yes/no questions and commands through different pitch excursions; and (5) affect and attitude of the speaker, such as emotional state, social persona, politeness strategies and attitude towards the listener (among other things), the latter being the main focus of the present research.

Many studies have demonstrated the role of f_0 in intonation for the aforementioned categories of information (Gussenhoven, Repp, Rietveld, Rump, & Terken, 1997; Knight, 2008; Ladd, 2008; Terken, 1991, 1994). The relationship between the voice source and information structure based on intonation, albeit less often touched upon in the literature, has also been explored, often in language-dependent contexts (Garellek, 2022). For example, creaky voice was found to be a common occurrence at the end of a phrase as a potential prominence and boundary-lending phenomenon for English (Redi & Shattuck-Hufnagel, 2001), and breathy voice was found at the beginning of read phrases of American English and Mexican Spanish (Garellek, 2014).

Evidence also suggests that voice source parameters contribute to the signaling of *focus* and *focal prominence*, showing changes locally in multiple source parameters not only in the focally accented syllables when compared to the neutral utterance, but also globally in terms of reductions to the rest of the utterance, attenuating the prominence of non-focal material. This was seen in both natural speech analyzed through inverse filtering (Ní Chasaide, Yanushevskaya, & Gobl, 2011; Yanushevskaya, Gobl, Kane, & Ní Chasaide, 2010) and in synthetic stimuli, with voice source manipulations done to a neutral utterance and presented to naïve listeners for evaluation (Murphy, Yanushevskaya, Ní Chasaide, & Gobl, 2018; Yanushevskaya, Murphy, Gobl, & Ní Chasaide, 2016). The RD parameter in particular appears to be especially effective for modelling prominence in synthetic speech even when f_0 is not varied (Murphy, Yanushevskaya, Ní Chasaide, & Gobl, 2019).

Stress is similar to focus at the phrasal level, where it can lend prominence to a word relative to the rest of the sentence (Ladd, 2008), but the term is used in a phonetic context instead of the more pragmatic meaning of focus, which is more concerned with putting emphasis on an element of particular importance for the speaker of an utterance. The primary function of stress is the lending of prominence to specific syllables in words, e.g., for pitch accent realization. Stress patterns in this context can affect the rhythm of a language and can also have a contrastive function in terms of differentiating between the meaning of words with the same spelling, depending on where the pitch accent falls (e.g., *subject* as a noun and *subject* as a verb, in English) (Arvaniti, 2020). Like focus, stress has been associated with a number of variable voice source parameters, e.g., f_0 , EE and UP in Irish (Ní Chasaide, Yanushevskaya, Kane, & Gobl, 2013), and increased glottal opening constriction was also found to correlate to pitch accents in American English vowels (Bird & Garellek, 2019).

Different dimensions of prosody are also used for different purposes depending on the language. Irish and English, for example, are intonation languages, in which pitch is used to convey meaning at a sentence level (in addition to some cases of lexical stress contrast mentioned above for English). Tone languages such as Mandarin Chinese, Thai, Hausa (Nigeria) and Mixtec (Mexico), on the other hand, use pitch variations to distinguish between words. Other languages such as Swedish and Japanese, while also categorized as intonation languages, also use pitch to distinguish between words in a more limited manner,

such as changing the meaning of a word depending on where the pitch accent (stressed syllable) falls (Nolan, 2006). Additionally, some languages also use voice quality to differentiate between vowels and consonants (segmental level) or have different voice qualities associated with specific tones (Gobl & Ní Chasaide, 2010).

The voice source has also been shown to play a role in sentence modes and declination. Declination, first introduced by Cohen and 't Hart (1982), is the gradual fall of pitch and other phonetic dimensions towards the end of a sentence. Building on previous research (Ní Chasaide et al., 2013), Ní Chasaide, Yanushevskaya, and Gobl (2015) have found indicators of declination of multiple voice source parameters across three sentence modes, which include declarative, wh- questions and yes/no questions, to different degrees. These parameter changes include declines in EE and RG, and increasingly lax phonation indicated by an updrift in OQ. Ní Chasaide et al. (2013) hypothesize that source parameters other than f_0 contribute to focal prominence to a greater extent whenever pitch prominence is not present, which their research seems to suggest. These parameters and their contributions, however, vary depending on the location of the accent and their sentence modes. Declarative and wh-questions showed a larger involvement of UP and EE, while yes/no questions relied mostly on RG and OQ. It bears mentioning, however, that there seems to be cross-speaker differences in how voice source parameters are used in focus and deaccentuation (Yanushevskaya, Ní Chasaide, & Gobl, 2017), and thus findings cannot be broadly generalized.

Larrouy-Maestri et al. (2024) define emotional prosody as “the mapping between acoustic information and emotions” (p.1). While a more thorough exploration of the literature on the acoustic parameters related to emotional states is presented in section 2.2.4, prosody and vocal expression of emotion are, as mentioned, deeply interconnected (Arvaniti, 2020). For example, a study by Liu, Rigoulot, and Pell (2015a) suggests that, when using non-sensical sentences to investigate affect perception in the voice, the prosodic rules of the listener’s language must be preserved due to listeners’ “cultural expectations” as “foreign sounding” material can influence their perception. Cowen, Laukka, Elenfeldt, Liu, and Keltner (2019) and van Rijn and Larrouy-Maestri (2023) also reiterate the relevance of key acoustic features in the communication of emotion through prosody, “highlighting the

complexity of the mapping between acoustic features and emotion in speech” (Larrouy-Maestri et al., 2024).

It is thus important to consider such prosodic dimensions in the study of vocal affect, as parameter levels related to affect expression may vary depending on the prosodic structure of utterances and of the language itself. In this thesis, it is hypothesized that the voice source parameter changes associated with different affective categories will be aligned to the prosodic structures of the analyzed utterances, i.e., that the most significant changes to voice source parameters will happen on the focal accents. Therefore, a summary on the prosody of Irish is presented below.

2.1.7. Irish Prosody and Intonation

A summary of the prosody of Irish is here presented, focusing on the overall intonational structure of the language and on the realization of pitch accents for declarative sentences of Irish, with the Kerry dialect being of particular interest.

Most studies on the intonation of Irish follow the Autosegmental-Metrical (AM) framework of intonational analysis (Gussenhoven, 1984; Ladd, 2008). These intonational analyses are primarily concerned with variations in pitch and carried out within the framework of the IViE project (Grabe, 2001), which is an adaptation of the ToBI system (Silverman et al., 1992) designed to account for variation within English dialects (Grabe, Nolan, & Farrar, 1998). While IViE includes multiple annotation tiers, the most relevant level for the present research is the phonological level, which is a formal linguistic representation on the utterance’s intonation pattern based on the presence of high (H) and low (L) tones that mark pitch accents in an intonational phrase. These tones are derived from the work of Pierrehumbert (1980) and Beckman and Pierrehumbert (1986), and the IViE labelling system is based on the work of Gussenhoven (1984) and Grabe (1998). A more detailed description of the AM model can be found in Cruttenden (1997) and Ladd (2008), and the IViE system is described in Grabe (2001).

The Irish dialects can be divided into three major groups, spoken in the language pockets of the Gaeltachtaí: Ulster (North), Connaught (West), and Munster (South). Connaught Irish encompasses a variety of smaller dialects, again broadly divided into South Connaught (Connemara and the Aran Islands) and North Connaught (Mayo), while Munster

Irish includes varieties spoken in Kerry and Cork (M. Dalton, 2008). The dialect of Irish explored in the experimental sections of this study is the dialect of the Dingle peninsula (Corca Dhuibhne) in West Kerry. A detailed account of Irish prosody can be found in Hickey (2011) and in M. Dalton (2008), and in O'Reilly (2014) for the Ulster and Connaught dialects. This summary will review Irish intonation in broader terms, without addressing intonational detail such as peak timing alignment or all the levels of analysis present in the IViE system. More in-depth views into Irish prosody can also be found in M. Dalton and Ní Chasaide (2003, 2005, 2007); Dorn and Ní Chasaide (2016); Dorn, O'Reilly, and Ní Chasaide (2011, 2012) and O'Reilly and Ní Chasaide (2015).

Irish is a stress-timed language, i.e., in which stressed syllables are articulated longer and louder than unstressed ones, which in turn become backgrounded (Dorn et al., 2012; Hickey, 2011). It has short and long vowels, both of which are consistently longer in stressed syllables than in unstressed syllables, and any stretch of Irish speech will contain a number of non-contiguous, acoustically prominent syllables that are regarded as stressed by speakers (Hickey, 2011, p. 306). Irish does not have contrastive intonational patterns to differentiate words, nor tone distinctions.

Word stress patterning varies between Irish dialects, with the most prominent differences being between the Northern/Western dialects and the Southern dialects (Hickey, 2011). While the former presents mostly word-initial stress (with some exceptions in Mayo Irish, which can feature stress on second syllables), the latter varies depending on syllable attraction (i.e., when stress patterns of a word can be influenced by surrounding syllables adjacent to the stressed syllable), making the stress patterns of Southern Irish sensitive to the prosody of the surrounding syllables. This results in Southern Irish having the greatest stress variation within the dialects, with increased phonological weight given to the long vowels (to which the Irish *fada*, i.e. the diacritic mark is added), even though stress can also be given to short vowels and certain consonant combinations.

Hickey (2011) also states that significant intonation contour differences were found between dialects, both for declarative sentences and other sentence types. Patterns in Donegal (Northern) Irish are clearly different from the Connaught and Munster variants. Varieties of Northern Irish feature a preference for a rising nuclear tone (M. Dalton & Ní Chasaide, 2003, 2005) across all sentence types explored in the literature (declaratives, wh-

and yes/no questions) (Hickey, 2011). This is also stated in M. Dalton (2008), who mentions that the predominant accent found across all sentence types of Donegal Irish are low rising tones labelled $L^* + H$ in the IViE system, and in M. Dalton and Ní Chasaide (2007), who state that this distinctive low rise is accompanied by a pitch plateau. On the other hand, Connaught and Munster Irish have more similar tonal inventories between them, and, in contrast to Donegal Irish, mostly feature falling accents. While the intonation patterns of these varieties have not been as thoroughly explored as Northern Irish, the literature does indicate that Kerry, Mayo and South Connaught Irish predominantly use a falling accent across declaratives, labelled as $H^* + L$, or a high falling accent (M. Dalton & Ní Chasaide, 2007). Additionally, Hickey (2011) also mentions that Donegal and Connaught Irish are said to have greater intonational movement through the sentence than the Southern dialects, and that these intonational patterns are also transferred to their regional Irish English counterparts.

These analyses of the typical intonation patterns of Kerry Irish in particular were used to inform the choice of recordings for analysis during the experimental stage of the present research, as it was desirable for utterances to match typical declarative sentence patterns of the dialect and avoid potential focus shifts.

2.2. Research on the Acoustics of Affect

Because of the overall complexity of the emotion process and of how to define emotion, research on the acoustic correlates of affect to date encompass a wide variety of frameworks and approaches for different stages. These can be broken down into the following categories, which are explored in this subsection: (1) research models on affect expression; (2) data collection methods (acted, induced or natural emotion), (3) experimental approaches for testing, and (4) data analysis which results in inferences on the acoustic correlates of emotion, i.e., measures based on mathematical models that may relate to certain affective categories in terms of global levels and contours.

2.2.1. Research Models on Vocal Expression of Affect

Many studies have provided evidence that listeners can correctly infer a speaker's emotional state based solely on the voice and its prosodic features such as pitch, loudness and speech rate (Frick, 1985) at accuracies much higher than chance. These include the

works and thorough reviews by Scherer (1986), Banse and Scherer (1996), Johnstone and Scherer (2000), Juslin and Laukka (2003), Scherer (2003), Juslin and Scherer (2005), Laukka (2008) and Laukkanen, Alku, Airas, and Waaramaa (2008). For example, early studies by van Bezooijen (1984) report an average rate of 65% of recognition accuracy across many different emotions. Similar results were found in Scherer, Banse, Wallbott, and Goldbeck (1991), and a number of early studies with emotion recognition accuracies of around 60% can be found in a review by Scherer (1989).

Scherer, Banse, and Wallbott (2001) have also suggested that, in adults, emotions can be communicated through speech alone. Juslin and Laukka (2003) reported the same for acted speech, with recognition accuracy rates approaching 70% for anger, fear, happiness, sadness, and tenderness in the voice. In a more recent study, Laukka et al. (2016) found that listeners were able to recognize most of the intended states from 11 categories of emotion in semantically neutral sentences. Studies with younger subjects have also shown that children and adolescents, albeit with less accuracy than adults, are also able to detect emotional information from affective speech prosody (Eskritt & Kang, 2003; Sauter, Panattoni, & Happé, 2013).

However, despite the apparent existence of acoustic correlates of emotion based on this human ability to infer emotion from voice-only stimuli, research into finding consistent sets of cues that deliver information about a speaker's emotional state still poses a challenge. These include methodological problems such as acquiring data and measuring voice source characteristics, as well as 'the existence of marked individual differences in the vocal expression of emotion' (Juslin & Scherer, 2005, p. 70). Such differences may arise from cultural and individual ways of expressing and perceiving emotion, as well as from the fact that the same acoustic result can be achieved through a variety of combinations of vocal settings or means in phonation and articulation of sounds (Kappas, Hess, & Scherer, 1991).

Another challenge is separating the acoustic features of a speaker's natural voice and the changes brought on by emotional arousal – in other words, the distinction between acoustic features resulting from speaker's personality (or long-term affective states) and natural, physical voice source characteristics, and those resulting from emotion (as defined by Scherer, 2003). Because of this, research frameworks have been developed to attempt to tackle this issue and guide affective speech research.

The Brunswikian Lens Model (Brunswik, 1956) and its theory of perception is widely used in affective speech research due to its incorporation of all stages of the communication process (i.e. articulatory, acoustic and auditory). The model is based on the Brunswikian Paradigm, which has as a main feature the connection between research methods and subject matter, considered appropriate for affective speech research due to its evolutionary perspective on human behavior, focusing on the relationship between the individual and the environment. The lens model separates the communication process into two parts: the encoding, or the production level (the externalization of a speaker's state); and the decoding, or the perception level (the attribution of the speaker's state by the listener, or the ability of judges to infer emotion from the voice) (Kappas et al., 1991). It has been used in many studies on vocal affect expression (Bänziger & Scherer, 2003; Grandjean, Bänziger, & Scherer, 2006; Juslin & Laukka, 2001).

Two other research models are considered by Scherer, Ladd, and Silverman (1984), Ladd, Silverman, Tolkmitt, Bergmann, and Scherer (1985) and subsequently by Bachorowski and Owren (2008) as the two main strands of research in affective speech acoustics, which investigate how affective states might interact with acoustic parameters and differ in terms of the types of cues to be considered: the *covariance model*, and the *configuration model*.

The covariance model, or continuous model, describes emotions based on continuous acoustic parameters such as those caused by physiological arousal, like fundamental frequency level and range, which vary continuously with arousal. The more intense the emotion, the higher the parameters rise. It is believed to better explain biologically based emotions, as information about the speaker's emotional state is encoded in the acoustic signal independent of linguistic meaning (Kappas et al., 1991). It may be said, however, that this covariation of features complicates the identification of acoustic properties that determine a listener's evaluation of the speaker's emotion, especially for the naturalistic expressions (i.e., not acted/induced for research), and therefore research that investigates the specific signal dimensions involved in the expression of emotion separately could potentially shed some light on the roles and impact of each feature (Arias, Rachman, Liuni, & Aucouturier, 2021).

The cue-configuration model, or categorical model, considers the possibility that different emotions are expressed through different combinations of acoustic cues, and

questions whether these parameters are reliable enough to measure emotional states (Juslin & Scherer, 2005; Scherer, 2003). It is also said to take into consideration linguistically relevant acoustic features: it suggests that categorical linguistic structures must be combined with verbal content (e.g., intonation changes in *wh*- questions in neutral affect utterances) and is more suitable for describing complex affective states that involve a more substantial cognitive component (Kappas et al., 1991). This approach is however questioned by the fact that no reliable set of cues has been found that effectively differentiates between emotions, despite studies proving that listeners are able to infer emotion from vocal-only stimuli at frequencies much higher than chance (Juslin & Laukka, 2001). This could be explained by the limited number of cues and approaches used by researchers in the field, and further investigation on cues less commonly used could potentially shed more light on the matter.

2.2.2. Speech Data Acquisition

A variety of methods have been used for the acquisition of data for affective speech research and are usually chosen based on accessibility and aims of the study. These methods can be divided into three categories: natural vocal expressions, induced stimuli, and acted vocal expressions.

Natural expressions may be acquired from databases of media recordings like, for example, the Belfast Naturalistic Emotion Database (Douglas-Cowie, Campbell, Cowie, & Roach, 2003), or by recording the every-day interactions of subjects over a long period of time, such as in the JST/CREST Expressive Speech Processing Project (e.g., Campbell, 2005; Campbell & Mokhtari, 2003). A study by Juslin et al. (2021) involved the use of tie-clip microphones that recorded the subject's speech for many days. While this provides ecological validity if one's objective is to study naturalistic emotions, there are a few methodological challenges related to the quality and the analysis of such stimuli. First, the recording quality is not uniform, as the positioning between the microphone and the speaker may vary and environmental noise might be included into the stimuli. Furthermore, this type of recording setup does not meet the requirements for inverse filtering and therefore an estimation of voice source characteristics, which is preferably done through recording in an anechoic chamber with highly specialized equipment (Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005; Ní Chasaide & Gobl, 2005). Another issue pointed out by

Scherer et al. (2003) and Juslin et al. (2021) regards the labelling of emotion categories, as not only are they self-reported by the speaker, but this process can also happen several days after the fact.

Induced stimuli involve, as the name suggests, inducing emotional states in the speaker to be recorded in the laboratory. This can be done through a variety of techniques, such as: suggesting mental imagery or asking speakers to retell emotional events (Harrigan, Lucic, & Rosenthal, 1991); showing emotion-inducing images and facial expressions (Ekman, Levenson, & Friesen, 1983); the Velten technique (Velten, 1968), in which the speaker reads emotional self-referent statements expressively and repeatedly (Barkhuysen, Krahmer, & Swerts, 2007; Wilting, Krahmer, & Swerts, 2006); experimenter-induced techniques in the classic social-psychological tradition, music, and hypnosis, though the latter is considered scientifically and ethically questionable (Juslin & Scherer, 2005). Electronic methods have also been used, including computerized emotion induction batteries that involved completing tasks under time pressure (Scherer et al., 2003) and computer games (Douglas-Cowie, Cowie, Cox, Amir, & Heylen, 2008; Zara, Maffiolo, Martin, & Devillers, 2007).

The induction approach has the advantage of being highly controllable while also providing the possibility to acquire natural, non-acted expressions with high ecological validity. However, each of these techniques was found to have differing degrees of effectiveness, oftentimes depending on the type of emotion (i.e., some were more effective for negative or positive emotions), and often resulting in relatively low intensity expressions (Scherer, 2003). As mentioned, ethical problems may also arise depending on the methodology used for the induction. The induced emotion approach has been used in studies by, for example, Aubergé and Cathiard (2003); Audibert, Aubergé, and Rilliard (2005); J. Barrett and Paus (2002); Batliner, Fischer, Huber, Spilker, and Nöth (2003); Batliner, Steidl, and Nöth (2008); Caridakis et al. (2006); Rossato, Audibert, and Aubergé (2004), and Zara et al. (2007).

Finally, acted emotions are acquired by using professional actors (or, at times, even non-actors) to portray the target emotion categories, often directed by researchers based on the aim of the studies and/or using either self-induction techniques (Abelin, 2004; Airas & Alku, 2004; Burkhardt et al., 2005; Sawamura et al., 2007), scenarios (Bänziger & Scherer, 2007; Scherer et al., 2001; Scherer et al., 1991; Schröder, 1999; Sobin & Alpert, 1999;

Williams & Stevens, 1972) or, at times, with the help of professional directors (Banse & Scherer, 1996; Bänziger & Scherer, 2007; Scherer et al., 2001).

The acted data approach has been criticized (Bachorowski & Owren, 2003; Russell et al., 2003) due to its lesser ecological validity, making it less generalizable for applications aimed for natural speech. Scherer (1986, 2003) states that acted portrayals of emotion provide much more intense and prototypical expressions than what is usually found in natural or induced emotional states, and “reflect sociocultural norms or expectations more than the physiological effects on the voice as they occur in natural conditions” (Scherer, 2003, p. 232) as a result of physiological arousal.

While differences in acoustic parameters between natural and acted emotions have been found (Audibert, Aubergé, & Rilliard, 2010; Jürgens, Hammerschmidt, & Fischer, 2011; Juslin, Laukka, & Bänziger, 2018; Scherer, 2013; Williams & Stevens, 1972), an argument can be made for acted portrayals when we consider the possibility that these same social norms and constraints regulate day-to-day interaction, and that public displays of emotion can be considered, to some extent, portrayals themselves (Banse & Scherer, 1996). This *ritualization* of emotional expressions, explored in Eibl-Eibesfeldt (1989) and in Shariff and Tracy (2011), suggests that human displays of emotion are not simple biological responses, but have rather become “more stereotyped and discrete for the sake of efficient and reliable communication” (Juslin et al., 2021, p. 2). The way in which humans express emotion is also constrained by the social environment and by the listeners themselves, depending on how the speaker wishes to be perceived. Accordingly, stereotypes used by actors are “expected to match those that listeners use in decoding the expression intention” (Scherer, 2003, p. 237), being therefore considered appropriate for perception studies.

As mentioned, Juslin, Laukka, et al. (2018) reported differences in the acoustic parameter measures between acted and spontaneous expressions of the same emotions; additionally, a follow-up study also revealed that listeners could distinguish between the two types of stimuli from arousal (Juslin, Harmat, & Laukka, 2018). Acted expressions tended to express more discrete emotions (possibly due to its stereotyped nature), while spontaneous expressions became more discrete when they presented higher intensity. This was also found by Sauter and Fischer (2018), and high-intensity clips in Juslin et al. (2021) also showed higher decoding accuracy for spontaneous expressions. Acted emotions have

also been perceived more strongly than induced emotions (Barkhuysen et al., 2007; Wilting et al., 2006).

The present study opted to utilize acted portrayals of emotion based on the aims of this research. Acquiring acted stimuli is more practical, avoids potential ethical issues and facilitates the acquisition of high-quality data based on the available hardware and recording conditions. Voice source analysis demands very specific quality requirements that would hardly be achievable in naturalistic settings, and many studies still rely on acted emotion. Additionally, considering the desired application of findings to speech synthesis, the emotional signal from acted portrayals of emotion seemed the most appropriate, as the simple modifications proposed in this study would not allow for the amount of detail and variation that might be present on naturalistic recordings.

2.2.3. Experimental Perception Approaches

In addition to mathematical models for measuring acoustic cues, theoretical and research models of emotion, and data acquisition methods, there are also many approaches on how to conduct perception studies on affect to acquire insight into which acoustic cues are related to different emotional states from a listener's perspective. Thorough reviews on perceptual approaches for the study of vocal communication of emotion can be found in Elfenbein and Ambady (2002), Johnstone and Scherer (2000), Juslin and Laukka (2001, 2003), Juslin and Scherer (2005), Kappas et al. (1991), Ní Chasaide and Gobl (2005), Scherer (2003) and Scherer et al. (2003). For the purposes of this project, the most widely used approaches are detailed in this subsection.

Encoding studies (or emotion recognition studies) are an approach with a long tradition and a large body of evidence produced. It evaluates the ability of listeners to infer the emotional state of the speaker, most often using acted portrayals of affect with semantically neutral utterances (i.e., sentences with no inherent emotional content) or pseudo-utterances (non-sensical sentences with made-up words that still match the prosodic and grammatical structure of the language of the study). These emotionally colored stimuli are presented to listeners, who are asked to indicate which emotion they perceive within a list of standard emotion labels (Scherer, 2003) and, sometimes, to rate the intensity of the emotion they perceive. Most studies utilize the 'Big Six' basic emotions (Ekman, 1999), although some scholars recommend the use of larger lists to account for a

wider range of emotions common in every day interactions (Cowie & Cornelius, 2003). These studies often find emotion recognition accuracy rates far above chance (Banse & Scherer, 1996; Scherer et al., 2003), the listener's answer being considered 'correct' if their response matches the target affect intended by the actor.

The use of labels is one of the major drawbacks of this approach, as the forced-choice task usually results in emotion discrimination rather than recognition, which essentially primes the listener to look for one of the given emotion categories in the responses while listening. This may inflate the accuracy rates based on a propensity of the listener to choose the available label that is closest to what was perceived (even if, in normal interactions, their decoding might have been slightly different or more nuanced) and even guessing, therefore requiring adjustments to the processing of results (Banse & Scherer, 1996; Elfenbein & Ambady, 2002; Elfenbein, Mandal, Ambady, Harizuka, & Kumar, 2002; Wagner, 1993). Some studies further opt to use additional labels such as 'no emotion' or 'none of the above', allowing listeners to label the emotion they perceive themselves (e.g., Abelin & Allwood, 2000; Greasley, Sherrard, & Waterman, 2000), but this can further add to the number of variables as free labeling is highly dependent on the listener's emotion vocabulary and, thus, harder to interpret.

Still, this method is the most widely used in the literature, as it is fairly straightforward for listeners and results are easier to analyze for researchers (Yanushevskaya, 2010). The results also do not seem to differ too much from other methods (Elfenbein et al., 2002). Early emotion recognition studies include van Bezooijen (1984), Scherer (1989) and Scherer et al. (1991).

The second most popular approach are *emotion inference studies*. While emotion recognition studies focus more on the decoding ability of the listener, this approach is more concerned with the nature of the underlying voice-emotion mechanisms – i.e., which specific cues are responsible for the emotion inferences of listeners (Scherer, 2003, p. 237). Instead of presenting purely recorded stimuli to participants, cue utilization studies systematically manipulate acoustic cues to evaluate which of these cues are responsible or more crucial for emotion recognition. These manipulations can be carried out in three different ways: (1) cue measurement and statistical association; (2) cue masking; and (3) cue manipulation via synthesis.

Cue measurement and statistical association involve measuring the acoustic characteristics of the emotion portrayals presented to listeners and correlate them through statistical means to the listener's judgements. The approach, however, is highly dependent on two factors: the nature of the stimuli used and whether they have sufficient variation in parameters, and the precision of these parameter extractions. Like the previous approach, listeners can judge the expressed emotion on multiple-choice lists, or, alternatively, scales (e.g. valence and intensity). For example, Bänziger and Scherer (2003) asked listeners to judge emotions on a number of scales (based on Granqvist, 1996) by moving cursors that corresponded to pitch (high – low), intensity (weak – strong), intonation (monotonous – modulated), speed (slow – fast), articulation (bad – good), instability (steady – shaky) and roughness (not rough-rough). Their results indicated that listener's assessment of perceived vocal features does provide insight into the vocal features responsible for the communication of emotion. Other studies using this approach include Banse and Scherer (1996), Scherer, Koivumaki, and Rosenthal (1972), Scherer, Rosenthal, and Koivumaki (1972) and van Bezooijen (1984).

In cue masking, certain acoustic cues are masked, manipulated, or entirely removed from recorded stimuli to investigate the effect of a particular cue on emotion inferences by listeners. This can be done through, for example, low-pass filters (e.g. in the early research of Alpert, Kurtzberg, & Friendhoff, 1963; Starkweather, 1956), or other masking techniques such as filtering, randomized splicing, backwards playing, pitch inversion, and tone-silent coding (Friend & Farrar, 1994; Scherer et al., 1984). Each technique will maintain part of the original speech, but remove or reduce other information, allowing researchers to isolate acoustic cues and identify which are the most crucial for affect inference, i.e., which type of cues carry the affective information. For example, Scherer et al. (1984) used random splicing to destroy the f_0 contour and semantic content but keep global f_0 characteristics and voice quality of utterances, their results suggesting that voice quality was the major carrier of affective information.

Cue manipulation via synthesis involves the systematic manipulation of vocal parameters to determine the effect of each type of acoustic cue on listener's perception of emotion. Early studies include Lieberman and Michaels (1962) and Scherer and Oshinsky (1977). The approach is further explored in Breitenstein, Van Lancker, and Daum (2001),

Burkhardt and Sendlmeier (2000), Cahn (1990), Carlson (1992), Carlson, Granström, and Nord (1992), Heuft, Portele, and Rauth (1996), Murray and Arnott (1993) and Murray, Arnott, and Rohwer (1996). Breitenstein et al. (2001), for example, compared original recordings of German sentences spoken in five different affects with the results from the same sentences with pitch and speech rate systematically altered, finding speech rate to be the strongest cue to emotion decoding. This approach was created with the aim of adding emotional coloring to speech synthesis systems (Scherer, 2003), which may also be done with resynthesis.

Copy synthesis, which is another term for resynthesis, also allows one to change different voice cues from a recorded neutral sentence by copying the characteristics of the recording through speech synthesis and systematically manipulating the digital sound wave. Some early studies used resynthesis to manipulate global f_0 , intensity and duration, and even accent structure (Ladd et al., 1985). In terms of voice quality, some early investigations into voice timbre manipulation were conducted in Burkhardt and Sendlmeier (2000), Carlson (1992) and Lavner, Gath, and Rosenhouse (2000) but, due to reasons mentioned previously, the voice source is much less explored than others acoustic dimensions. Nevertheless, this approach has provided evidence for the importance of voice quality as a carrier of affective information. Mozziconacci (1998), for example, used resynthesis of neutral utterances with modified emotional f_0 contours, finding that removing voice quality dramatically reduced recognition accuracy rates.

Audibert et al. (2005) used resynthesized French stimuli varying f_0 , intensity, duration, and voice quality at different intensities, results suggesting that no parameter alone was able to carry the entire emotional information. Furthermore, each cue seemed to signal different emotions at differing degrees: f_0 contours were more effective in expressing positive emotions, while voice quality and duration were best for expressing negative emotions, and intensity did not bring any significant information on its own. Burkhardt et al. (2006) used synthesized stimuli and manipulated pitch range, duration model and jitter simulation in a cross-cultural study, which showed both global as well as language-specific effects of each cue. Gobl and Ní Chasaide (2003b) used a formant synthesizer to produce stimuli with different voice qualities to investigate the relationship between the voice source and the expression of emotion, moods, and attitudes. Other

studies that use the cue utilization approach include Klasmeyer (1997), Protopapas and Lieberman (1997), Teshigawara (2004) and Waaramaa, Alku, and Laukkanen (2006).

In addition to these two main approaches, Scherer (2003) also mentions transmission studies and representation studies. Transmission studies focus mostly on the effects of transmission channels (e.g., telephones) on the voice and has been rarely used in the vocal communication of emotion. Representation studies involve measuring proximal cues based on listener's report of their impressions of a speaker's voice. This means that cues are not quite empirically measurable, as the approach relies on listener's reports of their own impressions of a speaker's voice and speech – which is constrained by the listener's vocabulary and language (in terms of emotion words and cross-language meanings), which often does not include research terms such as nasal, sharp, or rough (Scherer, 2003, p. 242). Impressionistic labelling is, however, partly what the work of Laver (1980) and other voice quality frameworks do when describing voice quality categories in terms of distal cues. Scherer (1982) has also designed rating sheets to systematize such judgements, but research using this approach has also been extremely rare (e.g., Scherer, 1974 for personality).

2.2.4. Acoustic Correlates of Emotion

Banse and Scherer (1996) have suggested that listeners use changes in cues such as pitch, loudness, timing, and voice quality to recognize the emotion of a speaker. Building on this and on previous studies, Juslin and Scherer (2005) divide voice cues most commonly used in affective speech research into four broad categories relating to: (1) fundamental frequency, including mean f_0 , f_0 range and variance and f_0 contours; (2) voice intensity; (3) voice quality, often based on the work of Laver (1980) and parameterized through models such as the LF model (Fant et al., 1985), and (4) temporal aspects of speech, including speech rate, articulation rate, and pauses.

Fundamental Frequency

Fundamental frequency (f_0) has been one of the most widely explored parameters in affective speech research, both in global and local terms (i.e., f_0 contours). Perceived pitch, which f_0 is the main determinant of, has been explored as a primary carrier of affective information in early studies such as Ohala's biological frequency code (Ohala, 1984) and

early studies by Fónagy (1978) and Huttar (1968), which examine how animals use lower pitch to assert dominance and size. Accordingly, a study by Noble and Xu (2011) investigating the acoustic correlates of friendliness found that ratings of “happy” and “friendly” were positively correlated to increases in pitch mean and range, particularly for happiness.

On the other hand, other studies have also found that the correlation between global f_0 and emotion is not as straightforward (Frick, 1985; Lieberman, 1968; Scherer, 1986). For example, Lieberman (1968) suggested that potency, rather than relating to lower pitch, results in increased loudness which causes an increase in pitch. An earlier review by Scherer (1981) suggested that anger, happiness, confidence and sometimes fear are expressed with increased pitch and increased loudness and speech rate, while contempt, boredom and grief have lower pitch and slower speech rate.

These contradictions may result from an issue with terminology or emotion categorization, such as frustration and aggression being both labelled as anger. However, Scherer (1986) suggests, given his own findings on the same emotion being expressed with both high and low f_0 , that these differences in f_0 levels between studies may be a result of psychological or physiological processes. Additionally, Scherer (1981) also addresses the issue that different emotions may present the same profile, suggesting that global f_0 levels alone are insufficient to accurately discriminate between the emotion categories used in his study.

The role of f_0 contours, or the variation of f_0 related to intonation has also been investigated, oftentimes in combination with the intensity and duration parameter (Fónagy & Magdics, 1963; Frick, 1985; Grichkovtsova, Morel, & Lacheret, 2012). Such studies indicate that f_0 contours play a strong role in signaling affect independently from linguistic content (Audibert & Aubergé, 2007; Carlson et al., 1992; Fónagy, 2001; Ladd et al., 1985; O'Connor & Arnold, 1973; O'Reilly & Ní Chasaide, 2007; Wells, 2006). O'Reilly and Ní Chasaide (2007) have suggested that “internal relationships within the utterance are likely to be as crucial as the global measures” (p. 10) that are most often the focus of research into characterizing intonational variations associated with certain emotions in terms of the timing of the nuclear f_0 peak in the syllable. Similarly, Fónagy (2001) also supports the idea of studying the dynamics of affective speech rather than utilizing mean values of acoustic parameters.

Ladd et al. (1985) reported on how contour shapes had greater than expected effect arousal scales, suggesting that uptrend intonation can signal arousal, annoyance and involvement. Carlson et al. (1992), using resynthesis of utterances and forced choice perception tests in order to study the effect of segmental articulation on perceived emotion, has suggested that local f_0 contour variation plays a significant role in the expression of affect. Grichkovtsova et al. (2012) found that prosodic contours (defined as a combination of f_0 , intensity and temporal parameters) that were extracted and transplanted between utterances with the same segmental content contributed to the identification of emotions to the same degree as voice quality measures taken for French, also playing a role in the identification of attitudes (as previously described by Juslin & Scherer, 2005). Additionally, the affective meaning of f_0 contours was also suggested to be dependent on the sentence type (e.g., question or statement) (Uldall, 1964). Wichmann (2000) and Bänziger, Morel, and Scherer (2003) also argue for the importance of context when inferring attitudinal meaning from prosodic contours.

In contrast, evidence against the correlation of emotion and f_0 contours also exists, such as the existence of the same f_0 contour profiles being found for different emotions (Mozziconacci, 1998, 2002; Scherer, 1981). The same studies found larger differentiation between emotions in terms of global f_0 levels such as mean and range (Bänziger et al., 2003; Bänziger & Scherer, 2005). Similarly, Bänziger and Scherer (2005) also found little evidence for quantitatively distinct contour shapes between different emotions, with, rather, global f_0 descriptors seeming to reflect the level of arousal of the portrayed emotions.

On the topic of global f_0 movement, Paeschke (2004) has found it to vary greatly relative to neutral utterances and considered it a suitable parameter to differentiate low activation emotions. For example, global trend seemed appropriate for the description of boredom and sadness, although almost meaningless for happiness and anger. Paeschke, Kienast, and Sendlmeier (1999) write that the basic emotions of fear, anger, happiness, sadness, boredom and disgust are successfully differentiated by f_0 mean and range, and also in the intensity of syllable stress, suggesting that emotions may have specific rhythms, and in the direction and steepness of f_0 contour changes. Ladd et al. (1985) has shown a large effect of f_0 range changes in judgements of arousal, and suggested that the f_0 contour is less significant than global f_0 level and voice when comparing studies between stimuli generated

by random splicing (destroying f_0 contour but keeping original voice quality characteristics) that yielded higher emotion recognition accuracy than stimuli generated through low pass filtering (keeping intonation contours but masking voice quality). Mozziconacci (1995) also found that highly recognizable emotions cannot be obtained by pitch manipulations alone, and that other features such as duration and voice quality must also be considered.

While f_0 has been often considered to successfully carry affective information to some degree, these inconsistencies between the studies mentioned above suggest that f_0 alone is not sufficient for creating distinct profiles for emotions or delivering the entire affective message. Scherer et al. (1984) and Ladd et al. (1985) argue that all the non-verbal aspects of f_0 range, contour and voice quality can carry the emotional content independently of linguistic content, but that recognition accuracy drops whenever voice quality cues are removed from the stimuli, indicating that the combination of f_0 , voice quality, intensity and time-related cues are necessary for affect recognition rather than one individual cue (Mozziconacci, 1998). Similar conclusions are drawn in Gobl and Ní Chasaide (2000, 2003b) and Ní Chasaide and Gobl (2002, 2005), where voice quality variations yielded higher recognition results than f_0 -only variations.

Voice Quality

Voice quality has also been a focus of studies on affective speech (Birkholz, Martin, Willmes, Kröger, & Neuschaefer-Rube, 2015; Gobl et al., 2002; Gobl & Ní Chasaide, 2003b; Grichkovtsova et al., 2012; Mozziconacci, 1998; Murphy, Yanushevskaya, Ní Chasaide, & Gobl, 2017; Murphy et al., 2022; Ní Chasaide et al., 2013; van Rijn & Larrouy-Maestri, 2023; Yanushevskaya, 2010; Yanushevskaya, Gobl, & Ní Chasaide, 2009; Yanushevskaya et al., 2018). Scherer (1986, p. 145) suggested that “although fundamental frequency parameters are undoubtedly important in the vocal expression of emotion, the key to vocal differentiation of discrete emotions seems to be voice quality, that is, the timbre of the voice, acoustically determined by the pattern of energy distribution in the spectrum”. This is further experimentally supported in Scherer et al. (1984). He also mentions that fictional literary works often describe emotion in characters’ voices as sounding tense, harsh or soft, among other descriptors often found in descriptions of voice quality. This suggests that voice quality, as commented by Yanushevskaya (2010, p. 10), “is a perceptual entity, and the relationship to the acoustic correlates is complex”. Despite this complexity, voice quality

has been shown to play a significant role in emotion recognition, with Scherer (2003) further commenting on the importance for perception studies to examine the psychoacoustics of affective speech such as perceived loudness, pitch and rhythm.

Voice quality was often neglected by early studies due to difficulties in measuring voice source characteristics, individual differences in natural vocal features to express emotions (physiological processes), and the variety of strategies speakers use to achieve certain acoustic results (Scherer, 1986). There has also been evidence of speakers using specific parameters in different ways to express emotion, such as differences in the degree of changes in f_0 declination and intensity (Aubergé & Cathiard, 2003). Still, the literature shows a significant amount of evidence for voice quality being one of the strongest carriers of affective information, with or without modifications in f_0 (Gobl et al., 2002; Mozziconacci, 1998; Ní Chasaide et al., 2013; Yanushevskaya et al., 2018).

Further support for the strong role of voice quality in the vocal expression of emotion can be found, for example, in Mozziconacci (1998), who found that there was a decrease in emotion recognition accuracy when voice quality cues were removed from a set of resynthesized utterances. In a study investigating the relationship between voice quality and affect across multiple cultures, Yanushevskaya et al. (2018) found that the stimuli that did incorporate non-modal voice quality, with or without f_0 variation, were generally more effective in cueing affect than those that only included f_0 variation. Other studies have also explored how voice source parameters, in combination with f_0 , map to affect in terms of both global, utterance-wide shifts in voice quality as well as local shifts aligned to the prosodic structure of the sentence, although results on which kinds of shifts are the most effective in cueing affect seem to vary depending on the emotion (e.g., Gobl et al., 2002; Gobl & Ní Chasaide, 2003b; Murphy et al., 2017, 2022; Yanushevskaya et al., 2009).

In contrast to f_0 , voice quality requires multiple covarying parameters to be successfully captured, as a single measure cannot fully capture all the characteristics of different voice qualities (Gobl & Ní Chasaide, 2003b; Juslin & Scherer, 2005; Monzo, Alías, Sanz, Gonzalvo, & Planet, 2007; Yanushevskaya, 2010). For example, Monzo et al. (2007) analyzed a wide range of voice quality parameters for differentiating between expressions of neutral voice, happiness, sadness, sensuality and aggressiveness in a Spanish corpus, and

concluded that all parameters used were important in differentiating each of the emotions. Patel et al. (2011) have found that 11 out of 12 parameters tested from set of acted short bursts had a significant effect on affect expression for relief, sadness, joy, panic fear and hot anger.

Still, even though differences in voice quality alone in modified synthetic stimuli were able to evoke different emotional colorings in an utterance to listener judges, Gobl and Ní Chasaide (2003b) found no “one-to-one mapping between voice quality and affect” (p.208), as some qualities appeared to be associated with more than a single affective state. For example, Ní Chasaide and Gobl (2005) found that a tense voice configuration correlates with perceptions of stress, anger, confidence and formality, while both lax and creaky states were connected to relaxed, bored and intimate affects. They believe that, while voice quality is critical in more subtle variations of emotional states, f_0 may still be the main parameter in the expression of stronger emotions. Because of that, the authors propose that affect be studied through an integrative approach that treats f_0 and voice quality together, while also considering them in relation to the temporal, rhythmic structure of utterances in order to integrate aspects of prosody related to its linguistic dimension and those that communicate speaker affect (Ní Chasaide & Gobl, 2004a, 2004b). They argue that this integration may be more effective in communicating affect than each of these features alone, and supporting evidence for the effectiveness of such an approach is also found in Yanushevskaya (2010), Audibert et al. (2005), and Audibert and Aubergé (2007). This is also echoed in a study by Grichkovtsova et al. (2012), where prosodic contour (f_0 , intensity and duration) and voice quality showed equally important contributions to the identification of emotions.

Yanushevskaya (2010) also explores the mapping of voice quality to affect by analyzing how f_0 perturbations contribute to voice quality features for emotion, concluding that, while voice quality by itself has shown higher recognition scores than f_0 alone, the most effective expression of emotion comes from combining the two sets of parameters. Audibert et al. (2005) conducted perception tests using resynthesized monosyllabic words from an acted speech database and varying combinations of f_0 , intensity, voice quality and time parameters, concluding that no single parameter carries the entire affective information. In their results, joy showed the highest recognition scores for f_0 only and f_0 and

intensity, while disgust, sadness and fear showed the highest scores for voice quality and duration. Their findings suggests that f_0 contours may reveal more information on positive emotions, as concluded also in Audibert and Aubergé (2007), while voice quality and duration may reveal more on negative emotions. They also comment that intensity changes did not reveal any important information on affective states when used alone, suggesting that the parameter is often complimentary on top of other cues of emotion. In a similar manner, Birkholz et al. (2015) found that while different phonation types had highly significant effects on emotion recognition rates, those were more crucial in the perception of fear, anger and neutral affect, while happiness, sadness, boredom and disgust seemed to primarily rely on other prosodic parameters (such as f_0) for German.

While these cue-based studies provide considerable insight into the relationship between voice quality and affect, as mentioned, thus far no definitive set of parameters has been determined for each emotional state. This is not only because of the reasons stated previously, but also because studies vary in terms of the emotion categories chosen to be investigated, as well as the models used for parameter extraction, which provide different acoustic measures and therefore different interpretations of results.

On the other hand, and albeit still presenting variations on the acoustic measures analyzed, studies following the covariation approach that have provided evidence for affect expression in terms of the dimensions of activation, valence and power can be found in Bachorowski (1999), Schröder et al. (2001), Russell et al. (2003) and Schröder (2004a). Schröder et al. (2001) concluded that most of the acoustic variables in their study correlated to emotional dimensions, especially activation, where active emotions showed “higher f_0 mean and range, longer phrases, shorter pauses, larger and faster f_0 rises and falls, increased intensity and a flatter spectral slope” (p.2). Other dimensions such as evaluation (valence) and power presented weaker correlations, but, as per the author, correlations nonetheless. Ryan, Ní Chasaide, and Gobl (2003) also did not find a positive correlation between happy voice and moderate tension in the tense-lax continuum as was suggested by Scherer (1986), but also suggest that perception of anger appears to be continuous in terms of the lax-tense continuum variation. A thorough review of how voice quality relates to affective states in many different ways can be found in Podesva and Callier (2015).

Intensity and Temporal Features

In addition to f_0 and voice quality, research also points to the importance of intensity and temporal parameters in the vocal expression of affect. Breitenstein et al. (2001) suggest that speech rate can provide an even larger contribution than f_0 variation to the expression of emotion, mentioning that sadness and fright showed a slower rate and anger a faster rate than the perceived neutral utterance. Morningstar, Dirks, and Huang (2017) report that happiness is typically linked to high pitch, high intensity and fast speech rate, while sadness is linked to low pitch, low intensity, and slow speech rate. These conclusions are also reflected in Johnstone and Scherer (2000), Juslin and Laukka (2001, 2003), Scherer (1996) and Pittam and Scherer (1993).

Other studies like Scherer, Koivumaki, et al. (1972), Scherer and Oshinsky (1977) and Bänziger, Hosoya, and Scherer (2015) also suggest that pitch, intensity and speech rate are important cues not only for affective information in the voice, but also for listener identification. Vroomen, Collier, and Mozziconacci (1993) have also found that rule-based manipulations to pitch and duration of a monotonous utterance was sufficient to express various emotions.

An early set of vocal parameters for 14 distinct categories of emotion can be found in Banse and Scherer (1996), which has been used in several subsequent studies (Aubergé & Cathiard, 2003; Bänziger & Scherer, 2003; Drioli, Tisato, Cosi, & Tesser, 2003; Rossato et al., 2004). Research in the area and acoustic profiles for different affect categories have also been reviewed and summarized in Frick (1985), Murray and Arnott (1993), Johnstone, Banse, and Scherer (1995), Johnstone and Scherer (2000), Klasmeyer and Sendlmeier (2000), Scherer (1986, 1989, 2000a, 2003), Scherer et al. (2003), Yanushevskaya (2010) and Kreiman and Sidtis (2011). Juslin and Laukka (2003) also provide a very thorough summary of studies and their respective investigated affects in the literature up to that point in time.

In summary, the same acoustic cues can be found in different emotions, or the same emotion may be expressed with different sets of cues, especially when considering global measurements. This is further complicated by the existence of different individual speaker baselines and different ways in which people perceive and express emotion (Aubergé & Cathiard, 2003). Additionally, there is also evidence that some parameters may carry more

weight for expressing a certain emotion than others (Audibert & Aubergé, 2007). The literature leads us to conclude that attempting to find a single, defined set of parameters for each emotion category even across a single language is a very daunting task. This study therefore suggests a more focused approach using a handful of voice source parameters to achieve acceptable vocal expression of emotion, while also considering the importance of the cultural context in the study of emotion in the voice.

2.2.5. The Voice Source and Prosodic Structures

Another area of importance to the current research is the relationship between prosodic structures (intonation) and voice source variation. While most studies in intonation focus on the role of f_0 (Gussenhoven et al., 1997; Hermes, 2006; Terken, 1991; Vainio & Järvikivi, 2006), evidence for the contribution of voice quality to the perception of prosodic structures such as pitch accents, focus and declination has also been found (Gobl & Ní Chasaide, 2010).

Ní Chasaide and Gobl (2004a, 2004b) argue that, because prosody is essentially the temporal patterning of the entire voice source (which includes more than just f_0), the dynamic patterning of the source (i.e., local voice quality modulations) is as important to the linguistic function of prosody as it is to the paralinguistic function of expressing emotion. In their research, they have found (1) declination shown by the EE parameter, (2) prominence of syllables being indicated by increased phonatory tension, and (3) phrase boundaries being typically characterized by increasingly breathy voice.

Iseli et al. (2006) also found positive correlations between voice source parameters and prosodic features like pitch accents, stress, and sentence types. Epstein (2003) has shown that prominence and phrase boundaries consistently affect voice quality in English, where prominent and phrase-initial words display a tenser voice quality than their counterparts. Similarly, Vainio, Airas, Järvikivi, and Alku (2010) investigate changes in voice quality related to prominence and focus in Finnish, showing results that contradict findings for English due to language-specific vocal strategies: in Finnish, speakers prioritize the use of f_0 over tenser voice quality and do not seem to modify the shape of the glottal pulse as would be expected considering other studies. Sluijter and Van Heuven (1996), Sluijter, Shattuck-Hufnagel, Stevens, and van Heuven (1995) and Sluijter, van Heuven, and Pacilly (1997) found that, for Dutch and American English, a tenser voice is associated with focal

stress and prominence. Thurgood (2004) reported that breathiness is related to emphasis in Javanese, while Callier (2013) found a relationship between creaky voice and the expression of interest and enthusiasm depending on its location in the sentence for Beijing Mandarin. Finally, Shiqing Zhang (2008) reported that voice quality in combination with other prosodic features can improve listener's recognition of emotion for Chinese natural speech.

This indicates that shifts in voice source parameters related to the prosodic structure of utterances can also vary depending on the language of the speaker, i.e., languages vary in terms of which and how different voice source parameters are used to convey certain prosodic structures such as focus and prominence, or even sentence types, as previously mentioned for Irish dialects (Hickey, 2011).

Many studies on Irish and Irish English have found variations in a number of LF model (Fant, 1995; Fant et al., 1985) voice source parameters correlating to focus (Ní Chasaide et al., 2011; Yanushevskaya et al., 2010), accentuation (Ní Chasaide et al., 2013), declination in voice source parameters (Ní Chasaide et al., 2015), f_0 declination, sentence mode and prominence (Murphy et al., 2018, 2019; Yanushevskaya et al., 2016) in terms of both global levels and local shifts.

Yanushevskaya et al. (2010) shows that accented syllables are not only marked by f_0 , but also by boosted EE, RG and UP. The same parameters showed lowered values for the non-focal material, along with higher OQ values, which suggest laxer phonation. This indicates that focally accented syllables involved increased respiratory effort, which causes a stronger excitation and a smaller open quotient. A subsequent study (Ní Chasaide et al., 2011) also found similar evidence, with the same parameter shifts that indicated tenser phonation found for accented syllables and laxer phonation for the non-focal realizations, but differences between sentences with rising and falling pitch: a rise in f_0 appeared to counteract the effects of deaccentuation. Similar results were also reported in Ní Chasaide et al. (2013) for Irish sentences, again with variations depending on sentence type.

In the case of Irish English, Yanushevskaya et al. (2016) found that the cueing of focus depends on the location of the focus in the intonational phrase, with significant differences between the non-final and final position in terms of voice source modulations

needed for cueing focal prominence. Similarly, in exploring the contribution of RD and correlated parameters EE and UP to the cueing of prominence, Murphy et al. (2018) suggest that the phrasal position of the intended focus makes a large difference to prominence judgements: post-attenuation of RD (i.e., laxer phonation applied to the part of the utterance following the accented syllable) was particularly important for cueing prominence in the first accented syllable (being equally or even more effective than boosting RD, i.e., adding tenser phonation to the syllable itself), but pre-attenuation (applied to the part of the utterance preceding the particular syllable) was more significant for the second accented syllable. This led to further exploration of how prominence could be modelled in synthetic speech through voice quality variation (Murphy et al., 2019), finding that both boosting of the accented syllable and attenuation of pre- and post-focal material contributed to the perception of salience, with no significant effect of focal syllable location on the magnitude of RD excursions within the same voice quality type. Additionally, on the topic of RD, Murphy et al. (2022) investigated the effects of global and local RD shifts on the expression of emotion in Irish in one of the few studies bringing Irish prosody and emotion together, finding that effective affect signaling does entail a combination of both global, utterance-wide voice source shifts as well as local shifts sensitive to the prosodic structure of the utterance.

Finally, in an investigation of voice source correlates for declination in Irish, Ní Chasaide et al. (2015) found that declination was indeed realized in source parameters and not only in f_0 , reporting a declination of EE, RG and OQ that indicated a gradual decrease in the level of source excitation, and an increasingly laxer phonation. Sentence types (declarative, wh-questions and yes/no questions) also showed subtle differentiation in phonation mode/source levels.

In short, there seems to be a fair amount of evidence for the relationship of the voice source and prosodic structures, given the importance of the voice source for the expression of emotion and the relationship between prosody and affect expression.

2.3. The Role of Context

While there is a lot of evidence for universal patterns in vocal communications of emotions, human interaction presents additional levels of complexity from the notion that

“humans are particularly cultural beings and, consequently, many behaviors that are rooted in biological systems can appear variably as a function of cultural evolutionary processes” (Bryant, 2021, p.26). Such behaviors may include the interplay between vocal communication and physical (face and body) expression, semantic context (given the evolution of language and meaning), and cumulative cultural evolution.

Cultural evolution also means that the ways in which we express or interpret emotion can be highly dependent on several factors. We may change our ways of expressing or even perceiving emotion depending on the person we are interacting with, and context effects like other sensory information, perceiver-based context and culture can strongly shape emotion perception (see review by L. F. Barrett, Mesquita, & Gendron, 2011). It would be sensible to say that the more information a listener receives on a speaker’s emotional state, the better their ability to accurately decode the emotion is.

This subsection explores the role of context in emotional expression and perception, with a review on the multimodal nature of emotion, the effects of facial expressions on vocal affect perception, the effects of the semantic context and, finally, cultural context and the specific aspects of individual speaker characteristics.

2.3.1. Emotion as Multimodal

As mentioned, many studies have provided evidence that listeners can correctly infer a speaker’s emotional state based solely on the voice and its prosodic features such as pitch, loudness and speech rate (Frick, 1985) at accuracies much higher than chance, seen in the work of Scherer (1986), Banse and Scherer (1996), Johnstone and Scherer (2000), Juslin and Laukka (2003), Scherer (2003), Juslin and Scherer (2005), Laukka (2008) and Laukkanen et al. (2008), among many others. Cowen, Elfenbein, et al. (2019) found that vocal bursts conveyed up to 24 distinct emotions bridged by gradients of continuously varying meaning. Emotion recognition accuracy from vocalizations and vocal bursts (such as cries, laughs, shrieks, and short vowel bursts) has also been explored in Scherer and Wallbott (1994), Sauter and Scott (2007), Simon-Thomas, Keltner, Sauter, Sinicropi-Yao, and Abramson (2009), Laukka et al. (2013), Cordaro, Keltner, Tshering, Wangchuk, and Flynn (2016) and Sauter et al. (2010), with results suggesting that emotions in vocal bursts are correctly recognized above chance levels in a number of globalized and remote cultures.

The literature also suggests, however, that the expression of emotion also affects multiple modalities of the body simultaneously, which are reflected in the measurable vocal parameters of each affect and play a complimentary role in the correct recognition of emotions by the listener. From an evolutionary approach, we may consider that the evolution of emotion signaling in humans evolved on a face-to-face basis, with long-distance communication technology being a very recent development in comparison. Because of that, it would be reasonable to suggest that many of the cues to emotion are presented in, for example, the visual channel, especially for more subtle emotions used in more familiar settings. This multimodality of emotion is widely accepted in the academic community, with a consensus that the origins of vocal expressions can be found in the physiological responses that accompany emotional responses (Juslin et al., 2021) and that the emotion-specific patterns of these changes create distinct acoustic patterns for the vocal expression of emotion (e.g., Scherer, 1986).

This is reiterated in a review by Keltner, Sauter, et al. (2019), who point to the importance of studying emotion as dynamic and multi-modal. They argue for the multimodality of emotion based on the assumption that “emotions enable people to react to significant stimuli (in the environment or internally), with complex patterns of behavior involving multiple modalities (...)” (id). They also mention evidence that emotions beyond the big six have “distinct multimodal and dynamic expressions” (p.138), and that those subtler, less basic emotions likely rely more on modalities apart from the voice. Indeed, several studies explore the effects of different expressive channels, often suggesting that different emotions rely more on cues from a specific channel than others for proper identification. These modalities are also mentioned in a review by Keltner et al. (2016), who explore theoretical developments on the aforementioned shaping of emotion perception based on context, culture and individual differences.

Looking into every one of these modalities in detail would warrant extensive discussion. For the purposes of the present research, this review will focus on the visual channel and the cultural context, briefly touching upon the semantic channel and individual speaker characteristics, and on studies connecting these modalities to the vocal expression of emotion.

2.3.2. The Visual Channel

In day-to-day interaction, vocal expressions of emotion are usually accompanied by the visual channel, i.e., facial expressions, which provides additional cues of emotion and improves affect recognition (Scherer, 1986). The study of facial expressions has a long tradition, starting with Darwin (1998) and following the same theoretical approaches to emotion as detailed previously. Thorough reviews on facial expression research can be found in Keltner and Ekman (2000) and Keltner, Sauter, et al. (2019); for the purposes of the present research, this review will focus on studies that bring together the vocal and visual channels for emotion expression and perception.

Juslin and Scherer (2005) advocate for the importance of studying the facial channel alongside the vocal channel, as they are “are similar in that both are characterized by a constrained physical structure, around which inter- and intra-individual variations convey information about the person’s identity and emotional state (...)” (p.66). Indeed, the effects of facial expression on voice perception is well documented (Borod et al., 2000; De Gelder, 2000; de Gelder & Vroomen, 2000; Massaro, 2000; Scherer, 2003; Scherer & Scherer, 2011), and studies also suggest that some emotions are better recognized from the face, and others from the voice (Scherer, 2003). Different recognition scores found between emotions in faces and voices are also mentioned in a review by Ekman (1994). For example, joy has been found to be unmistakably recognized in the face, but with a much lower recognition rating from the voice, while vocal expressions of sadness, anger and fear are usually well identified (Scherer, 2003). Johnstone and Scherer (2000) suggest that these differences come from an evolutionary pressure to make vocal expression of emotion clearer to the listener by adding the visual channel, which is in line with speculation by Juslin et al. (2021) that facial expressions may help listeners identify mixed everyday-life emotions when voice-only stimuli do not achieve the higher accuracy levels scored for emotions usually identified as a single discrete category.

A few types of visual stimuli are used in this area of study, such as: still photographs or extracted video frames of discrete emotional states (Borod et al., 2000; Paulmann, Jessen, & Kotz, 2009; Pell, 2002), sometimes including multiple versions/intensities of each emotion (Chronaki, Hardwin, Garner, Maurage, & Sonuga-Barke, 2014); sets of still pictures portraying continuums (de Gelder & Vroomen, 2000); animations/moving synthetic faces

(Massaro & Egan, 1996); or even recorded video of speakers (Paulmann & Pell, 2011). Stimuli are usually acquired from pre-validated corpora (e.g., Ekman & Friesen, 1975), or validated beforehand by the researchers conducting the main study.

These visual stimuli may be presented separately to investigate recognition accuracy differences (Chronaki et al., 2014; Pell, 2002), or combined with the vocal stimuli in different ways: they may be integrated (i.e., presented together at the same time), either congruently (each channel presenting the same target affect, e.g., as in de Gelder and Vroomen (2000) and Paulmann and Pell (2011)) or incongruently (channels presenting conflicting target affects, e.g., in Borod et al. (2000) and Massaro and Egan (1996)); or one may precede the other, the former serving as a primer for the latter (e.g., Paulmann & Pell, 2010). These studies often investigate the effects of each individual channel as well, comparing the emotion accuracy of uni-modal, bi-modal or, in some cases, multi-modal stimuli (adding, for example, the semantic channel or other additional source of information such as in Borod et al. (2000); Paulmann et al. (2009); and Paulmann and Pell (2011)). Listeners may also be instructed to focus on one channel over the other when making emotion judgements to investigate the influence of each channel (de Gelder & Vroomen, 2000).

Overall, the literature strongly indicates that the combination of vocal and visual stimuli, when presented congruently, consistently improves emotion recognition accuracy, with facial expressions being generally easier to recognize than voice-only stimuli (Chronaki et al., 2014; Pell, 2002). Investigating both channels separately, Pell (2002) found significant differences in emotion recognition between channels for the Big Six emotions, with speaker gender also playing a role in the recognition of disgust and the neutral expression. Recent studies also indicate that vocal tract characteristics are influenced by emotional orofacial gestures such as expressions of disgust (Chong, Kim, & Davis, 2018) and smiling (Arias, Belin, & Aucouturier, 2018), with the facial expressions themselves actively contributing to the vocal output of emotion.

However, mixed results arise when studies investigate which of the two channels is dominant in emotion recognition (e.g., Paulmann & Pell, 2010; Rigoulot & Pell, 2014). Collignon et al. (2008) found that congruent combinations of the visual and vocal channel were better recognized than each channel presented separately; for incongruent combinations, on the other hand, subjects showed a bias toward the visual channel when

attributing emotion. In contrast, de Gelder and Vroomen (2000) suggest bi-directional links between the two channels: bimodal presentation and judgement of stimuli appears to lead listeners to combine the two sources of information, but, when instructing subjects to ignore one channel over the other, their judgements were still influenced by the modality they were told to ignore. Rigoulot and Pell (2014) conclude that vocal cues of emotion facilitate the recognition of emotion in facial expressions and serve to guide a listener's visual attention to the most relevant cues of a given emotion. Paulmann and Pell (2010) also suggest that a "brief exposure to prosodic cues can establish a meaningful emotional context that influences related facial processing" (p.230), but more so when the prosodic information is longer in duration.

In contrast, Kraus (2017) questions the primary role of the face in the communication of emotion, finding that vocal cues, in combination with semantic cues, are more critical to emotion recognition than facial cues. He suggests that emotional expression across multiple modalities divides the viewer/listener's attention, and that voice-only communication with all its constituents is the most powerful channel for expressing emotion due to the interaction of linguistic and paralinguistic features (i.e., how something is said) and speech content (i.e., what is said).

Other factors such as gender and age also appear to affect emotion recognition in the face and voice. For example, Chronaki et al. (2014) suggest that younger listeners recognize vocal expression of emotion less accurately than adults, while recognition of facial expressions presents a lesser degree of difference between groups. This is consistent with a study by Morton and Trehub (2001) which showed that, when the linguistic and paralinguistic cues in speech conflicted, children seemed to rely more on the linguistic content compared to a higher reliance on prosody from the adult subjects. Chronaki et al. (2014) also found age-related differences in emotion attribution, with preschoolers having a higher tendency to attribute happiness to voices and faces and more accurately recognizing anger and happiness, while the recognition of sadness and other low-intensity emotions seemingly followed a slower developmental trajectory, being better recognized by comparatively older subjects (also in Widen & Russell, 2008).

In terms of gender, findings are inconsistent. McClure (2000), Collignon et al. (2010) and Kapitanović, Tokić, and Šimić (2022) showed slight advantages for female observers in

emotion recognition for both the voice and the visual channel, with the magnitude of gender differences depending on the specific emotion, the gender of the model, and other factors (see meta-analysis by A. E. Thompson & Voyer, 2014); however, other studies such as Sauter et al. (2013) and Chronaki et al. (2014) found no gender differences in the recognition of vocal emotion, although the ages of the listeners between these studies differed.

Several studies also indicate the existence of cultural differences in visual and auditory stimuli processing. Tanaka et al. (2010) suggest that Japanese listeners are more attuned to vocal stimuli than Dutch listeners, while Liu, Rigoulot, and Pell (2015b) report that American English speakers were more susceptible to the visual stimuli in contrast to Chinese listeners. Such differences have also been found when analyzing the semantic context in combination with vocal emotion expressions, as outlined below.

2.3.3. The Semantic Context

To validate the choice made for the present study in utilizing semantically neutral utterances, a small review of the effects of the semantic context in the perception of emotion is presented. Overall, studies suggest that the semantic context also improves emotion recognition in the voice when the combined stimuli are congruent (Borod et al., 2000; Lin, Ding, & Zhang, 2020; Paulmann et al., 2009; Paulmann & Pell, 2011; Pell, Jaywant, Monetta, & Kotz, 2011). The semantic context is usually incorporated as simply the linguistic content of the vocal stimulus (e.g., Paulmann et al., 2009; Pell et al., 2011), produced with affective or neutral prosody, with few studies that also investigate the unimodal condition presenting the semantic channel in text form (e.g., Borod et al., 2000; Paulmann & Pell, 2011) as sentences or emotion-denoting words (e.g., Cowie et al., 1999).

Results regarding channel dominance are mixed. Kotz and Paulmann (2007) and Paulmann and Kotz (2008) suggest that semantic content overrides emotional prosody, and that prosody can be ignored when it contrasts with the semantic context. However, their research also suggests that prosody and semantics interact in an additive manner (Beaucousin et al., 2007; Nygaard & Queen, 2008; Wambacq & Jerger, 2004). There is also evidence for cultural differences regarding dominance of the semantic, prosodic or visual channels in emotion recognition (Ben-David, Multani, Shakuf, Rudzicz, & Lieshout, 2016;

Ishii, Reyes, & Kitayama, 2003; S. K. Kim & Sumner, 2017; Kitayama & Ishii, 2002; Lin et al., 2020; Nygaard & Lunders, 2002).

It is important to note that each of the studies mentioned in this and previous sections use different methodologies, varying in terms of emotion categories, number of emotions tested, culture of speakers and listeners, and testing instructions. Therefore, while there is evidence for cultural differences in the effects of different channels in vocal affect recognition, methods may also be the cause of some of the reported differences. Still, the clear effect of the semantic channel in the perception of vocal affect is what led the present study to choose sentences with neutral semantic content for data collection.

2.3.4. The Cultural Context

It has been widely suggested that the context of an interaction plays a significant role in the accuracy of emotion identification. Kappas et al. (1991) state that “communication is not context free but is embedded in a framework of cultural or social rules and information, shared by the participating individuals” (p. 203). The idea of human expression of emotion being social is also mentioned in Ekman and Cordaro (2011), who state that “our emotional repertoire consists of fundamentally different, but socially constructed, ways by which we react to universally shared stimuli” (p. 364). While many similarities are known to arise given the shared evolutionary history of humans, there are also many reports on the differences of emotion decoding across different populations.

Elfenbein and Ambady (2002) suggest that even some moderate degree of consistency across widely different human groups indicate the existence of a shared mapping between acoustics and affect. Many studies have reported that emotions are recognized by listeners through prosodic cues even between speakers of different languages and cultures (Koeda et al., 2013; Laukka et al., 2016; Pell, Monetta, Paulmann, & Kotz, 2009; Scherer et al., 2001; W. F. Thompson & Balkwill, 2006; Waaramaa, 2014; Waaramaa & Leisiö, 2013, among others), sometimes also combined with facial expressions (Bryant & Barrett, 2008) albeit with different degrees of cross-cultural recognition, with some emotions being more accurately perceived in cross-cultural experiments than others (Beier & Zautra, 1972; Keltner & Haidt, 1999).

Other studies report mixed results, with some emotions being recognized cross-culturally, and others only recognized within cultural boundaries (Cordaro et al., 2016; Laukka et al., 2013; Sauter et al., 2010; W. F. Thompson & Balkwill, 2006) or in terms of affective properties like valence and arousal (Gendron, Roberson, van der Vyver, & Barrett, 2014a). Yanushevskaya (2010) and Yanushevskaya et al. (2018) found that different cultures perceive certain voice qualities as related to different categories of affect. Yanushevskaya, Ní Chasaide, and Gobl (2011) report both universal and culture-specific trends in the association of voice quality to affect; Kitayama, Mesquita, and Karasawa (2006) found that joy was expressed through different cues in Japanese and in French; and studies such as Cheang and Pell (2009), Ní Chasaide and Gobl (2004a), and Elfenbein and Ambady (2002) also found differences between different parameters and voice qualities.

Podesva and Callier (2015) consider these differences as evidence that voice qualities have culture-dependent social meaning and can be purposely controlled by the speaker rather than being simply a physiological response to the emotion being felt. Keltner, Sauter, et al. (2019) mention that even dialects of a single language may differ in how individuals express emotion (also reported in Cordaro et al., 2018; Elfenbein, Beaupré, Lévesque, & Hess, 2007; Laukka et al., 2016). Laukka, Neiberg, and Elfenbein (2014) reported high recognition accuracy of emotion between five different English-speaking cultures, but with the existence of in-group advantages, similarly to a study by Elfenbein (2013). Laukka et al. (2016) further confirms the existence of culturally specific schemas for emotion expression and perception, and more evidence is given in Esling (1978), Stuart-Smith (1999), Trudgill (1974), and Pell et al. (2009). Crivelli, Jarillo, Russell, and Fernandez-Dols (2016) also suggest the existence of possible cultural differences in how emotions are labelled.

The universality debate also applies to facial expressions. Izard (1994) argues for the innateness and universality of a limited set of facial expressions of emotion; in an early study, Ekman, Sorenson, and Friesen (1969) also found agreement in emotion identification in photographs from both literate and preliterate cultures, suggesting some pan-cultural element in facial displays of emotion. However, differences in what evokes the emotion, how the emotion display is controlled, and the consequences of expression were shown to differ. On the other hand, many studies have also reported cultural differences on facial

recognition of emotion across cultures (Gendron, Roberson, van der Vyver, & Barrett, 2014b; Haidt & Keltner, 1999; Jack, Blais, Scheepers, Schyns, & Caldara, 2009; Jack, Garrod, Yu, Caldara, & Schyns, 2012; Marsh, Effenbein, & Ambady, 2003; Tanaka et al., 2010; Yuki, Maddux, & Masuda, 2007).

This brings back the argument that expression of emotion may follow socially constructed, culture-specific display rules. Matsumoto, Yoo, and Fontaine (2008) explored cultural differences in expression of emotion accounting for overall expressivity norms, endorsement, and in- and out-of-group contexts and mention how a more prominent number of individuals with personality traits such as extraversion (or a higher social value put onto such traits) may change the way in which a culture perceives emotion. A review by Pisanski et al. (2016) explores how humans modulate their voices depending on different social contexts and speaker motivation, and the ability of humans to produce convincing facsimiles of an intended affect in order to acquire the benefits from expressing that affect has also been documented (e.g., Bryant & Aktipis, 2014).

Cordaro et al. (2016, p. 119) state that emotion expression recognition rates may vary as a function of culture for at least three reasons: cultural variations in emotion display rules, which can influence how often and how intensely members of that culture expresses a particular emotion; differences in which emotions are more valued in day-to-day interaction, which may yield variations in recognition and produce different gradients of such expression (Tsai, Knutson, & Fung, 2006; Tsai, Louie, Chen, & Uchida, 2007); and, finally, culture and language-specific prosody, which may result in culturally specific ways of expressing certain emotions (Marsh et al., 2003). This assumption is in line with the general literature on the topic of cross-cultural expression of emotion in the voice and reiterates the importance of investigating the acoustic correlates of emotion within specific cultural contexts.

2.3.5. Individual Speaker Characteristics

Juslin and Scherer (2005, p. 70) state that “the existence of marked individual differences in the vocal expression of emotion” is yet another challenge in affective speech research and attempts to create definitive sets of parameters for emotion categories. Such differences arise not only from the cultural and individual ways of expressing and perceiving emotion, but also from the fact that the same acoustic result can be achieved through a

variety of combinations of vocal settings or means in phonation and articulation of sounds (Kappas et al., 1991). For example, a study by Yanushevskaya et al. (2017) revealed considerable differences in the realization of accentuation and de-accentuation in terms of voice source parameters when comparing six different Irish English speakers, concluding that “the extent to which different source parameters and f_0 contribute to focal salience does appear to differ from speaker to speaker” (p.1037).

Podesva and Callier (2015) also suggest that voice quality changes can indicate individual speaker characteristics in terms of anatomy of the vocal apparatus and its development in terms of vocal behavior, and may also vary according to linguistic structures in languages to contrast sounds, or language tendencies (i.e., creaky voice at the end of phrases in English (Henton & Bladon, 1988) and Finnish (Ogden, 2001)). For example, voice quality can vary in terms of community patterns (Henton & Bladon, 1985), cultural gender conditioning (Hejrná, Šturm, Tylečková, & Bořil, 2021), expressing social personas (Callier, 2010; K. Hall, 1995; Podesva, 2007) or social status (Esling, 1978; J. A. Hall, Coats, & LeBeau, 2005; Knowles, 1978; Kraus, Park, & Tan, 2017; Stuart-Smith, 1999), cultural identity (Mendoza-Denton, 2011; Szakay, 2012), and language background (Mennen, Scobbie, de Leeuw, Schaeffler, & Schaeffler, 2010).

Yoonjeong Lee, Keating, and Kreiman (2019), albeit focusing on speaker recognition, also report speaker variation in several vocal parameters and suggest that familiarity with a speaker increases emotion recognition by the listener, arguing for the existence of person-specific vocal patterns. On the other hand, awareness of their vocal profile can also lead speakers to develop different vocal strategies to convey their emotional message more efficiently depending on the listener. They mention that because social context and affective state both cause variability in the voice (in both psychological and physiological terms), no speaker says the same utterance in the exact same way twice, especially considering that the voice can be consciously manipulated (e.g., see review by Kreiman & Sidtis, 2011).

Considering all this and given the intended future application of findings from the present research, it was decided that the production data used for this study would be collected from the same speaker as one of the voices of the ABAIR synthesizer, which could

facilitate implementation down the line. It bears mentioning that this was only made possible thanks to the speaker's availability and acting background.

2.4. Affective Speech Synthesis

In the past few decades, we have seen the development of many methods of producing speech through artificial means, with significant recent improvements to the naturalness of speech synthesis. Yet, it is still not quite able to capture and reproduce the expressiveness of human speech in its entirety. Given how beneficial advancements in that direction would be to those who use speech synthesis to express themselves, much effort has been put into adding more expressiveness and control to synthesis systems.

In essence, there are two options to add emotion to synthetic voice: direct synthesis, where the transformation is done from text straight into the target emotion before reconstruction of the speech output signal (i.e., without the generation of a neutral output) (Govind & Prasanna, 2013; Klatt, 1980, 1987); and voice conversion, where the emotion of the synthesized speech is adapted through an additional module after the synthesis of the neutral, non-expressive voice (Arias et al., 2021; Mohammadi & Kain, 2017; Stylianou, 2009). While most text-to-speech systems perform voice conversion, some go directly from phoneme sequences to acoustic structures (Triantafyllopoulos et al., 2023). Most studies and architectures follow the discrete/ categorical approach to emotion.

This section covers speech synthesis methods both in terms of direct synthesis and voice conversion, focusing on studies that have explored the implementation of affect to the synthesis, followed by a brief overview on the current methods used for Irish speech synthesis.

2.4.1. Direct Emotional Speech Synthesis

Synthesis approaches discussed in this section can be split into two categories: rule-based approaches, which rely on the principles of speech production and generates speech through models based on such principles, e.g., articulatory (Birkholz, 2013; Rubin, Baer, & Mermelstein, 1981) and formant synthesis (Cahn, 1990; Klatt, 1980, 1987; Murray, 1989; Murray & Arnott, 1995); and corpus-based approaches, which take natural recordings and use the analysis and parametrization of the waveforms to digitally model speech through

statistical frameworks, e.g., concatenative synthesis (Hunt & Black, 1996; Moulines & Charpentier, 1990), SPSS (King, 2011; Zen, Tokuda, & Black, 2009) and machine learning methods (J. Shen et al., 2018; Sotelo et al., 2017; van den Oord et al., 2016; Y. Wang et al., 2017).

An early review of methods on the simulation of emotion in synthetic speech is presented by Murray et al. (1996). Updated reviews are also provided by Schröder (2001, 2009), Govind and Prasanna (2013), Sisman, Yamagishi, King, and Li (2020), Triantafyllopoulos et al. (2023), and Barakat, Turk, and Demiroglu (2024).

Articulatory and Formant Synthesis

Most early approaches to general speech synthesis utilized methods such as articulatory, formant, and concatenative synthesis. Articulatory synthesis is mainly concerned with modelling the articulatory processes of speech, which is done by analyzing articulatory data acquired by measuring the speech organs and using such knowledge to create speech models that generate realistic sounds. Most articulatory synthesizers use the source-filter theory of speech production (Fant, 1960/1971) as a basis for the speech synthesis from the acoustic parameters. While this method should, in theory, produce highly natural-sounding synthetic speech, data acquisition is very challenging, and the construction of the model and mapping is very computationally complex (Govind & Prasanna, 2013). While it has not yet been used for expressive synthesis, advances to articulatory synthesizers continue to be made, a more recent example being the Vocal Tract Lab software (Birkholz, 2013).

Formant synthesis is also developed from the source-filter theory but simulates the vocal tract resonances (formants) and acoustic source signals of each sound rather than using a digital simulation of the vocal tract physicality. One of the most popular formant synthesizers was created by Klatt (1980). The extended version of this synthesizer, KLSYN88 (Klatt & Klatt, 1990), incorporates a version of the LF glottal flow model (Fant et al., 1985).

The first attempts to add emotion to the synthetic voices were done through rule-based approaches guided by speech experts and evaluated through listening experiments (Triantafyllopoulos et al., 2023). One of the earliest examples of expressive speech synthesis through formant synthesis is that of Cahn (1990) and the development of the Affect Editor,

which uses an acoustic listener model for manipulating perceptual parameters including pitch, timing, voice quality and articulation and uses DECtalk3 as its formant synthesizer (Conroy, Vitale, & Klatt, 1986).

The HAMLET system (Murray, 1989; Murray & Arnott, 1995) is another system that uses formant synthesis through DECtalk. In contrast to the Affect Editor, it processes its input by rule, without the need for it to be manually annotated. Other examples include the data-driven approach of Burkhardt and Sendlmeier (2000) using the KLSYN88 formant synthesizer (Klatt & Klatt, 1990), and a rule-based approach by Reddy, Harikrishna, Rao, and Manjunath (2015).

Concatenative synthesis and SSPS

Concatenative synthesis involves joining segments of natural speech from a database, and works on the principle of unit selection (Hunt & Black, 1996). Signal processing techniques such as linear prediction or the Pitch Synchronous Overlap Add (PSOLA) algorithm (Moulines & Charpentier, 1990) can be used to smooth the transitions between units.

For expressive synthesis, the unit database is “expanded to include the sound of each phoneme when expressed with each of the targeted emotions” (Crumpton & Bethel, 2015, pp. 275-276) and seen, for example, in Iida, Campbell, Iga, Higuchi, and Yasumura (2000); Iida, Campbell, and Yasumura (1998); Yilmazyildiz, Mattheyses, Patsis, and Verhelst (2006) and Fernandez and Ramabhadran (2007). This method can also be combined with copy synthesis (Heuft et al., 1996; Vroomen et al., 1993) or prosody rules (Mozziconacci, 1998; Rank & Pirker, 1998).

More recently, another widely used synthesis method is Statistical Parametric Speech Synthesis (SPSS), a corpus- and model-based approach that uses a vocoder for the analysis and synthesis of speech signals (King, 2011; Zen et al., 2009). These statistical models have been used to add expressiveness to synthetic voices, and include early machine learning approaches such as Hidden Markov Models (Kawahara, Masuda-Katsuse, & de Cheveigné, 1999; Pitrelli et al., 2006; L. Qin, Ling, Wu, Zhang, & Wang, 2006; Y. Qin, Zhang, & Ying, 2010; Tokuda et al., 2013; Z. Wu et al., 2014) and Gaussian Mixture Models (Aihara, Takashima, Takiguchi, & Ariki, 2012; Jia, Zhang, Meng, Wang, & Cai, 2011; Shen

Zhang, Xu, Jia, & Cai, 2008). Popular SPSS systems used for analysis, modification and resynthesis are the STRAIGHT system (Kawahara, 2006) and the WORLD Vocoder (Morise, Yokomori, & Ozawa, 2016). HMM-based systems were also used for transformation of speaker styles and emotion (Yamagishi, Masuko, & Kobayashi, 2004), as well as for speaker adaptation (Yamagishi et al., 2009).

Deep Learning and Neural Networks

Advanced machine learning approaches involving deep neural networks have been at the forefront of the most recent advances in affective speech synthesis (Tits, 2019). Examples include Wavenet (van den Oord et al., 2016), Char2Wav (Sotelo et al., 2017), Tacotron (Y. Wang et al., 2017) and Tacotron 2 (J. Shen et al., 2018). These approaches can be divided into two broad categories: supervised approaches and unsupervised approaches (see Barakat et al., 2024 for a thorough review). Supervised approaches for expressive synthesis involve models trained on datasets with emotional labels (i.e., happy, angry, sad; or even speaking styles), and most earlier models were supervised (Inoue, Hara, Abe, Hojo, & Iijima, 2017, 2021; Suzié, Nosek, Sečujski, Pekar, & Delié, 2019; P. Wu et al., 2019; Xue, Zhu, An, & Xie, 2018).

Unsupervised approaches automatically extract the target style or prosody, and synthesize similar speech based on that extracted information. This may be through direct reference encoding (Skerry-Ryan et al., 2018; Y. Wang et al., 2017), used to generate emotional speech in Younggun Lee, Rabiee, and Lee (2017) and Ren et al. (2019); using latent features via variational auto-encoders (VAEs) (Akuzawa, Iwasawa, & Matsuo, 2018; Kingma & Welling, 2014); global style tokens (GST) (Valle, Li, Prenger, & Catanzaro, 2020; Y. Wang et al., 2018); and in-context learning based approaches (Le et al., 2023; K. Shen et al., 2023; C. Wang et al., 2023).

While these approaches have been relatively successful at adding expressiveness to synthesis and producing very natural-sounding speech, they require large amounts of data for training. There is also the issue of controllability over the expressive output, be that at the global level (utterances, paragraphs, sentences) or at the local, more fine-grained level (word, syllables, phonemes). Some systems allow for some degree of controllability, but the controllable parameters depend on the architecture of the model. For example, STRAIGHT

(Kawahara, 2006) and WORLD (Morise et al., 2016) allow some degree of control over f_0 , spectral, cepstral and aperiodicity features, while others focus on spectral features (e.g., Younggun Lee et al., 2017; Zhou, Sisman, Schuller, & Li, 2023).

The first kind of parametric control mentioned in a review by Barakat et al. (2024) is through modeling-specific prosody rules (Hono et al., 2020; Younggun Lee & Kim, 2019; Nakata et al., 2022). The Controllable Expressive Speech Synthesis (ConEx) by Cornille, Wang, and Bekker (2022), for example, proposes prosody control at global (utterance) and local (phone) levels, outputting an integrated approach between the two levels.

Another method involves modeling-specific prosody features, such as f_0 /pitch, duration, energy, noise and spectral tilt (Barakat et al., 2024). Again, some studies propose control over global, utterance levels (K. Lee, Park, & Kim, 2021; Raitio, Li, & Seshadri, 2022), while others explore local levels such as phonemes, syllables and words (Gong et al., 2021; C. Kim, Um, Yoon, & Kang, 2022). Although this kind of fine-grained control is highly desirable for prosodic subtleties and speaking styles, achieving such specific prosody targets still proves a challenge.

Prosody modeling through strength focuses on controlling the intensity of the emotion or target prosody by having the model learn different emotion strengths through a variety of ML techniques (Choi & Hahn, 2021; Im, Lee, Kim, & Lee, 2022; Lei, Yang, Wang, & Xie, 2022; Li, Wang, Xie, Wang, & Xie, 2022; Lorenzo-Trueba et al., 2018; Yoon et al., 2022). Finally, prosody clustering uses clustering methods to select appropriate prosody embeddings for the model (Um et al., 2020), keeping phonetic and prosodic information separate during training for the sake of a fine-grained level of prosodic control (Vioni et al., 2021).

2.4.2. Emotional Voice Conversion

Another method to apply emotional changes to synthetic voices is through voice conversion. In general, voice conversion is defined as a way of making the speech of a source speaker sound like it came from another speaker, while keeping the linguistic content intact by changing all the speech parameters available in a specific model. In the context of affective speech synthesis, the technique was adapted to change the emotion of a base sentence while keeping the content and the speaker the same. This differs from direct

speech synthesis in the sense that modifications are applied after the base utterance is generated as an extra step in the synthesis pipeline, rather than applying the emotional parameters to the text input before the reconstruction of the signal. Because voice conversion relies on acoustic features already synthesized by the TTS system, this additional step splits the process into synthesis and conversion, which reduces computational complexity and data dependence and makes it the approach used by most TTS architectures (Triantafyllopoulos et al., 2023).

Thorough reviews of voice conversion methods can be found in Arias et al. (2021), Mohammadi and Kain (2017) and Sisman et al. (2020), the latter with a particular focus on Deep Learning approaches. For the purposes of this thesis, a brief overview on the basic concepts of voice conversion is presented, focusing on emotional voice conversion.

Voice Conversion Methods and Speech Research

Some emotional voice conversion approaches use direct transformation (utterance to utterance), while others use decomposition (i.e., decomposing an utterance into smaller components, with target emotion/style being one of them). Arias et al. (2021) mention that one of the main contributions of voice conversion approaches to the scientific field is the ability to change singular acoustic cues while keeping the all the rest of the speech signal the same, including other acoustic and contextual features. This allows researchers to examine listener's reactions to that specific type of cue change, investigate the effects of each of those cues on the perception of affect and formulate expert knowledge on the acoustic cues of affect categories.

Technically speaking, transformations can be done to the voice source (glottal transformations, as explored below), or to the filter (vocal tract) (Arias et al., 2020; Narendranath, Murthy, Rajendran, & Yegnanarayana, 1995; Quené, Semin, & Foroni, 2012), each with a few different approaches to different parameters. Global transformations used in the literature include fundamental frequency (f_0), intensity, and glottal pulse periodicity patterns.

F_0 can be changed via simple algorithms like, for example, the DAVID toolbox (Rachman et al., 2018), based on a technique that introduces small delays to an audio signal and raises/lowers the pitch by playing it faster or slower (Dattorro, 1997). This, however,

affects filtering, adding artifacts to the output. More recent techniques involve adjusting mean pitch level and pitch variability (Pell & Kotz, 2011), allowing for the creation of different pitch contours (Bachorowski & Owren, 1995; Jiang & Pell, 2017). This can be implemented through the PSOLA algorithm (Moulines & Charpentier, 1990) in Praat (Boersma & Weenink, 2002, 2024); through a phase-vocoder (Moulines & Laroche, 1995) such as that implemented in CLEESE (Burred, Ponsot, Goupil, Liuni, & Aucouturier, 2019); or through pitch-adaptative analysis techniques as seen in STRAIGHT (Kawahara, 1997). Most of these pitch shifting methods can also manipulate duration and speech rate of utterances (Arias et al., 2021, p. 15) if deemed necessary, depending on the acoustic correlates of the target emotion (Scherer & Oshinsky, 1977).

The intensity of the speech signal can also be modified in CLEESE (Burred et al., 2019; Ilie & Thompson, 2006). Finally, measures such as jitter, shimmer and harmonic-to-noise ratio, which involve changes in the shape and periodicity of the glottal pulses considered important for some emotional expressions (Gobl & Ní Chasaide, 2010; Johnstone & Scherer, 1999) can also be changed (Boersma & Weenink, 2002).

Arias et al. (2021) mention, however, that not all glottal source changes can be easily simulated with voice transformations. One reason for this is that machine learning models rely on specific sets of training data to determine pulse variability (e.g., Bóhm, Audibert, Shattuck-Hufnagel, Németh, & Aubergé, 2008; Bonada & Blaauw, 2013), and it is unclear how the patterns learned from these examples can be applied to voices not present in the dataset. Moreover, because glottal source estimation is computationally complex and slow, analysis systems have not yet reached the level of quality needed for robust automatic glottal source analysis, and manual, expert analysis takes time and is not feasible for large amounts of data, as previously mentioned. However, Arias et al. (2021) also suggest that synthesis approaches which involve direct manipulation of the voice source signal may provide better experimental control than voice transformation techniques.

Machine Learning and Deep Learning Approaches

Like synthesis, the most recent methods for voice and emotion conversion have used data-driven approaches like Deep Learning, encompassing several speech processing

techniques in their pipelines, from analysis to mapping and to reconstruction (Sisman et al., 2020).

Speech analysis and reconstruction can be achieved through signal-based representations like PSOLA to modify f_0 (Moulines & Charpentier, 1990) or the Harmonic plus Noise model (HNM) (Stylianou, 2001). Alternatively, model-based representations based on, for example, the source-filter model, are the basis of parametric vocoders like STRAIGHT (Kawahara, 2006) and WORLD (Morise et al., 2016). STRAIGHT is widely used for voice conversion precisely because its parametric nature allows for easy speech manipulation. Additionally, deep learning alternatives to parametric vocoders also exist (Goodfellow et al., 2014; van den Oord et al., 2016; Yamamoto, Song, & Kim, 2020).

The intermediate step between analysis and reconstruction, i.e., feature mapping, tends to be the most challenging part for machine learning models, and are usually divided between parallel (e.g., Abe, Nakamura, Shikano, & Kuwabara, 1990; Shikano, Nakamura, & Abe, 1991) and non-parallel training data approaches (e.g., Erro, Moreno, & Bonafonte, 2010). Parallel data approaches for affective synthesis include Aihara et al. (2012), who applied some changes to voice quality; Tao, Kang, and Li (2006), who synthesized three different emotion intensities in terms of variations to f_0 , durations and intensity; and Türk and Schröder (2010), who modified voice quality, f_0 and duration of a neutral TTS German utterance to target *cheerful*, *depressed* and *aggressive* affective states. For unparallel data, studies exploring emotional voice conversion include, for example, Luo, Chen, Takiguchi, and Ariki (2017), who attempted to capture f_0 features of emotions from micro-prosodic to sentence levels with a less than sufficient emotional data for training, and Y.-C. Wu, Hayashi, Tobing, Kobayashi, and Toda (2019), who used a DNN-based vocoder that allows for some degree of control over parameters.

The nature of Deep Learning approaches for voice conversion has, however, the same issues and limitations regarding its applicability to research on vocal correlates of emotion as the synthesis approaches (Arias et al., 2021). While some provide a small degree of parametric control needed for scientific experimentation, it is still far from the desired fine-grained controllability and extraction of information that would facilitate the identification of factors that may characterize emotion expressions, and it is therefore not as scientifically informed as more manual methods. Neural networks also do not allow for

easy introspection of the information representation within the network, meaning that it does not inform the researcher about which or how vocal features are being manipulated, limiting its interpretability.

While some degree of success at adding affect to the synthetic voice through machine learning has been achieved, research into emotional speech synthesis is still needed, especially considering cultural differences and the development of multi-language and multi-speaker expressive TTS applications. This point is further explored by Tits, El Haddad, and Dutoit (2019), who mention that the need for large amounts of data by current deep learning algorithms is still a big challenge for low-resource languages.

The difficulties related to voice source analysis and modelling also mean that the dimension of voice quality is rarely ever explored within machine learning approaches, even though it has been shown to be a major carrier of affective information and an important element in terms of cultural specificities in the expression of emotion.

There is a strong tendency in the field of expressive speech synthesis to focus on deep learning algorithms and improve their performance, even with smaller amounts of data. That prospect, however, may be far off into the future for low-resource languages with very limited number of speakers, especially one such as Irish with significant prosodic variation between dialects. This issue, combined with the current limitations of expressiveness control in synthesis and voice conversion systems, is what motivated the manual nature of the present study and the potential future application of findings to Irish assistive technology and language learning applications.

2.4.3. Speech Synthesis and Affect Conversion for Irish

This section explores the current available speech resources and technologies for the Irish language, including corpora, speech models, and synthesis tools.

A list of the available spoken speech corpora for Irish can be found in Chapter 3 of the Digital Plan for the Irish Language (Department of Tourism, 2023, p. 37). While there has been a considerable degree of effort into acquiring more audio data and making corpora available as open resources, the collection of Irish language corpora is still recent, especially compared to other languages. While the current amount of data for Irish is suitable for data-

driven and statistically based approaches such as machine learning and deep neural networks for text-to-speech synthesis and automatic speech recognition, it still needs considerable expansion to match the level of performance achieved for major languages with large corpora availability.

In terms of speech models, the only parametric model used to capture the tone-of-voice dimension for Irish at the time of writing has been the LF model (Fant, 1995; Fant et al., 1985) within the Róbóglór project, explored in previous studies conducted at Trinity College Dublin's Phonetics and Speech laboratory (Gobl & Ní Chasaide, 2003b; Murphy, Yanushevskaya, Chasaide, & Gobl, 2021; Murphy et al., 2018; Ní Chasaide et al., 2015; Yanushevskaya et al., 2018; Yanushevskaya et al., 2017). These studies encompass prosodic dimensions such as focus and declination as well as the expression of affect.

The ABAIR project at Trinity College Dublin has been involved in developing speech corpora, linguistic resources and speech technologies for the Irish language (Murphy et al., 2023). Synthetic voices have been developed for the three main dialects of Irish, and these and other resources are made available to the public (www.abair.ie, Ní Chasaide et al., 2017). A variety of synthesis systems have also been developed through the course of the ABAIR project. Irish synthetic voices, but based on L2 speakers of Irish, have also been developed by *Microsoft* (female, non-Gaeltacht) and by the Scottish company *Cereproc* (female Ulster, non-Gaeltacht) (Department of Tourism, 2023).

The ABAIR Irish synthetic voices are currently being used in a variety of assistive and language learning applications (Barnes et al., 2023; Barnes et al., 2022; Ní Chasaide et al., 2023; Ní Chasaide, Barnes et al., 2022; Ní Chiaráin & Ní Chasaide, 2019, 2020; Zuckermann et al., 2021). One such application is Geabaire (Barnes et al., 2023), an Irish language AAC system, which allows non-speaking people to communicate through the selection of pictures and words on a screen, which are concatenated and then spoken out via synthetic speech. In this application, where the synthetic voice is the user's own voice, it is particularly desirable to allow the user some means of communicating the emotion behind the spoken utterances, and this may best illustrate the context for the current explorations.

The two voice conversion systems available and relevant to the proposed research consist of model-based approaches involving the LF model, and development has been

focused on academic research and investigation of voice source correlates of emotion and prosody, as they allow for precise manipulation of some isolated parameters.

GlórCáil (Murphy, 2021; Murphy et al., 2020) is a manual analysis-and-synthesis system that allows for the manipulation of the voice source parameters f_0 , EE and RD. It has been used in several experiments conducted at the Phonetics and Speech Laboratory to explore the relationship between voice source, prominence and affect implemented in the MATLAB environment (The Mathworks Inc., 2022). It carries out the inverse filtering frame-by-frame using the modified version of Iterative Adaptive Inverse Filtering (Alku, 1992) GFM-IAIF (Perrotin & McLoughlin, 2019), providing an estimation of the differentiated glottal flow for voice source parametrization, along with filter coefficients that describe the vocal tract transfer function. Source parametrization is performed on each glottal pulse through the method described in Kane, Yanushevskaya, Ní Chasaide, and Gobl (2012), generating a set of vocal tract and voice source parameters by matching the differentiated glottal flow to the LF model (Fant et al., 1985) through a method based on dynamic programming (Kane & Gobl, 2013).

The synthesis interface takes these parameters as input and resynthesizes the speech using concatenated pulses from the LF Model instead of the original glottal source signal, which are then filtered by an approximate representation of the vocal tract transfer function. During the resynthesis stage, the utterance can also be manually manipulated in a mouse-controlled environment that visually presents the f_0 , EE and RD contours. It also allows for vocal tract filter transformation, i.e., length manipulation. Some of the aims of the system were to transform speaker characteristics to create new voices and manipulate prosody in terms of focus and affect for use in language learning and assistive applications.

In a later development, the GlórCáil system was also integrated into a deep neural network (DNN) based SPSS TTS framework (Murphy et al., 2021), still allowing for the manipulation of the aforementioned parameters in the output and aiming for the generation of almost any utterance (Murphy, 2021). The WORLD vocoder (Morise et al., 2016) is also used here to extract $\log f_0$, spectral and aperiodicity parameters, and to resynthesize the speech (Murphy et al., 2021).

The GloRí software (J. Dalton, Kane, Yanushevskaya, Ní Chasaide, & Gobl, 2014) is a glottal research system developed before GlórCáil that served as both its precursor and basis. GlóRí is a speech analysis interface that offers two approaches for voice source analysis: a manually optimized analysis, which locates GCIs automatically and allows the user to conduct the inverse filtering by setting formant frequencies and bandwidths, then fitting it to the LF model for parametrization; and a fully automatic analysis, with many parameter options derived from the algorithm-based analysis of the inverse filtered glottal signal. The system also presents a synthesis and parameter modification interface similar to GlórCáil, but using RG and RK instead of RD.

The Voice Source Generator (VSG), also developed at the Phonetics and Speech Laboratory (Z. Wang & Gobl, 2023b), came as a recent alternative to GlórCáil for synthesis and voice source modifications. It is a system for generating voice source signals, based on the LF Model (Fant, 1995) and designed with the MATLAB App Designer (The Mathworks Inc., 2022). Currently, it does not have an automatic analysis system, rather taking as input LF parameter data either from GlórCáil or from manual analysis software. The VSG allows for manipulation of the f_0 , EE and RD parameters, and for the user to input specific time points and respective target values.

The system was developed alongside the present research and used in the experiments described in this thesis after extensive testing of all available tools (further described in the methodology section). The methodology for using the VSG for this project is also explored in more detail in chapter 4.

2.5. Summary

A summary of the decisions made for the present study based on the existing literature explored in this chapter is presented here.

The definition of emotion used in this research follows that of Juslin and Scherer (2005), as short but intense spontaneous reactions to events that cause physiological arousal. This definition is widely used in the field of vocal affect expression and aligns with the aims of the present research in terms of what the proposed shifts are meant to provide for users of TTS systems in future implementation (i.e., the ability to express some degree

of vocal emotion when utilizing the synthesis system). The model of emotion research used is the discrete approach, which aligns with the objective of creating a set of simple modifications for a small number of discrete emotion labels. This is a traditional approach in the field and aligns with the methodology of data collection that is currently the most appropriate for detailed voice source analysis (described in Gobl & Ní Chasaide, 2010), namely, discrete emotional samples of acted data collected in a controlled, quiet environment with high quality recording equipment.

A review of voice quality frameworks was presented to introduce the concept of tense/lax voice and vocal fold tension, which is one of the modifications explored in the experimental chapters. The inclusion of creak (and other voice quality categories) was also motivated by considerations arising from the experiments' results, given the association seen in the literature between certain affects with voice qualities not explored in this study (e.g., Yanushevskaya et al., 2018). The methodology used for analysis of the voice source was manual inverse filtering, chosen due to its superior quality over automatic voice source estimation methods. The parametrization model chosen was the Liljencrants-Fant model (Fant, 1995; Fant et al., 1985), as it has been widely used in the field of affective prosody and on the research that this study builds upon (e.g., Gobl & Ní Chasaide, 2003b; Murphy et al., 2022), shown good performance in creating resynthesized stimuli, and its parameters allow for straight-forward calculation of the vocal tension parameter RD.

Given the proposed investigation of voice source parameters related to affect in terms of global (utterance-wide) and local shifts aligned to the prosodic structure of utterances, a review on intonation and prosody was also presented. It was followed by an overview on Irish prosody, detailing the typical prosodic structure of Irish dialects (Hickey, 2011) and informing the subsequent culling of data. This was to ensure that sentences chosen for analysis were declaratives with only the target affective coloring affecting its prosody, avoiding focus shifts and other prosodic structures that were not intended for exploration in this study and that could affect global and local measures.

This research experimentally follows the covariance model, where parameters are modified in global and local terms, aiming to find parameter values for each modification type to successfully cue the target affects. However, the possibility of different modification types (and information channels) being more effective for cueing different affects is also

considered. The present study utilizes copy synthesis (as seen in Gobl & Ní Chasaide, 2003b; Ladd et al., 1985) in order to create the stimuli presented to listeners, but does not seek to investigate which of the modified parameters is the most effective in cueing emotion categories; rather, it focuses on the effectiveness of modification types (local, global or combined), and on the effects of the addition of the visual channel to the perception of vocal affect.

A summary of findings on the acoustics of affect was also presented in order to further inform decisions made during the resynthesis of affective stimuli, as well as to provide evidence that singular parameters applied globally to utterances are not enough to convey the full affective message of a speaker, i.e., there is no one-to-one mapping between acoustic parameters and affect categories (Gobl & Ní Chasaide, 2003b). While duration measures were not explored in this study due to software limitations, f_0 , intensity (in the form of EE) and voice quality (RD) were, covering three of the four main categories used in affective speech research (Juslin & Scherer, 2005). An overview on the connection between the voice source and prosody was also presented, as it has been argued that the temporal patterning of the voice source is crucial in the realization of linguistic prosody (Ní Chasaide & Gobl, 2004a, 2004b), and that prosody is deeply interconnected with affect expression (Arvaniti, 2020).

This research was kept within the context of the Irish language, given the evidence for culture-specific trends in vocal emotion expression and recognition. Data was collected from a native speaker of Irish, and perception tests were presented primarily in Irish to avoid any issues with cross-language emotion labelling. Informants were all Irish speakers.

The visual stimuli were included in the present research in the form of animated 3D models to specifically investigate the effects of this channel in the vocal expression of emotion for Irish, while also being motivated by the intended future applications of findings given the inherent presence of facial expression of TTS users during in-person interactions. The recorded utterances described in chapter 3 were purposely created to be semantically neutral due to the documented effects of the semantic channel on vocal emotion perception (Paulmann & Pell, 2011; Pell et al., 2011). Unfortunately, no studies specifically on facial expressions were conducted within the context of Irish, and so a more generalized methodology was used for the creation of the visual stimuli (described in section 7.1).

Despite reported differences in emotion expression between dialects (Cordaro et al., 2018; Keltner, Sauter, et al., 2019) (for languages other than Irish, as no comparative studies on this topic have been made for Irish as of the time of publication), it was decided not to exclusively seek listeners that were also speakers of the Kerry dialect due to considerable limitations in the number of available informants. The experiments were therefore conducted considering the wider context of the Irish language, even if the creation of stimuli focused on the Kerry dialect.

Individual differences between speakers reported in the literature (Juslin & Scherer, 2005; Yanushevskaya et al., 2017) in prosodic realization were the main reason why it was decided to record the same speaker that was used in the development of the ABAIR synthesizer, along with availability and acting skills.

As for synthesis, the manual approach described was developed because, at the time of writing, there is not enough data on affective Irish speech to consider training a Neural Network or other data-driven synthesis approaches. On a positive note, this allows for a more detailed investigation of voice source parameters related to emotion and more control over the synthesis stage, which aligns with the goal of creating voice source modifications that can be consistently applied to a synthesizer output.

In this study, an alternative approach to voice conversion is used. The baselines used are resynthesized versions of the neutral recording (experiment 1) or the ABAIR synthesizer (Chasaide et al., 2017) neutral baseline (experiments 2-4) based on the manual voice source analysis, and modifications are applied to the analysis parameters to compensate for the fact that, as of the time of writing, ABAIR has no voice source information that can be directly modified in the synthesizer output. F_0 is changed through the PSOLA algorithm (Moulines & Laroche, 1995), and other parameters are resynthesized through the LF model using the VSG system (Z. Wang & Gobl, 2023b), considered preferable to other systems due to its performance and user-friendly interface. Further detail on the methodology of this study is presented in chapters 3 and 4 below.

3. METHODOLOGY

This chapter details the methodology of the current research, aimed to address the research questions previously presented. To explore the effects of global and local changes in voice parameters on the perception of affect in the voice, methods of data acquisition, analysis and parametrization via the LF model (Fant, 1995) are presented, followed by the approach used to develop a set of systematic changes to the voice source parameters and its application to a neutral baseline. Finally, the approach used for conducting the perception experiments and analysis of results is outlined.

3.1. Production Data Acquisition

The data acquisition approach chosen for the present research was that of acted data, which provides better experimental control over the acquired stimuli as well as the high-quality recordings necessary for voice source analysis. The data recorded for the analytical phase of this study consists of eight Irish sentences recorded in a semi-anechoic chamber at the Phonetics and Speech Laboratory, using professional high quality Brüel & Kjær 4191 microphone with 2669-C preamplifier and Nexus 2690 conditioning amplifier to ensure phase linearity (which is required for voice source analysis) through the Audacity software (Audacity Team, 2021). Speech was recorded at a 44100 Hz sampling rate and saved as multiple .wav files.

The recorded subject is an adult male speaker of the Kerry (Dingle) dialect of Irish and a professional actor, whose voice also features in the ABAIR project synthesizer (Ní Chasaide et al., 2017). His voice was chosen due to potential ease of applicability to the ABAIR synthesizer in the future.

The elicitation materials were created loosely based on passages from Irish children's stories currently available on LARA (Zuckermann et al., 2021). Segmental phonetic considerations were prioritized over dialect-specific morphosyntactic and lexical preferences. This was due to a need of certain vowel characteristics for optimization of voice source analysis, such as vowel openness (with less obstructions in the vocal tract during production) and length (to better avoid the influence of stops and other adjacent sounds).

Table 1 - List of recorded Irish sentences for the present research with respective phonetic transcriptions.

ORIGINAL SENTENCE, GLOSS & PHONETIC TRANSCRIPTION		ENGLISH TRANSLATION
Sentence 1	Beidh sé a' báistí ar an lá sin. WILL IT RAIN ON THE DAY THAT. bʲɛ ʃɛ bʲɑːʃtʲiː ɛɹʲ ənʲ lʲɑː ʃɪnʲ	It will rain on that day.
Sentence 2	Gheobhmid bád ar an lá sin. WE WILL GET BOAT ON THE DAY THAT. ʃomʲɪdʲ bʲɑːdʲ ɛɹʲ ənʲ lʲɑː ʃɪnʲ	We will get a boat on that day.
Sentence 3	Beidh sé a' báistí amárach. WILL BE IT RAIN TOMORROW. bʲɛʃ ʃɛ bʲɑːʃtʲiː əmʲɑːɹʲəʁkʰ	It will rain tomorrow.
Sentence 4	Beidh Barra ar an bád leo. WILL BE BARRA ON THE BOAT WITH THEM. bʲɛʃ ʃɛ bʲarʲa ɛɹʲənʲ bʲɑːdʲ lʲoː	Barra will be on the boat with them.
Sentence 5	Beidh bád ann amárach. WILL BE BOAT THERE TOMORROW. bʲɛʃ ʃɛ bʲɑːdʲ aunʲ əmʲɑːɹʲəʁkʰ	There will be a boat tomorrow.
Sentence 6	Beidh Barra ar an bád leo amárach. WILL BE BARRA ON THE BOAT WITH THEM TOMORROW. bʲɛʃ ʃɛ bʲarʲa ɛɹʲənʲ bʲɑːdʲ lʲoː əmʲɑːɹʲəʁkʰ	Barra will be on the boat with them tomorrow.
Sentence 7	Beidh an aimsear go hálainn amárach. WILL BE THE WEATHER BEAUTIFUL TOMORROW. bʲɛʃ ənʲ aɪmʲʃɪɹʲ gə hɑːlʲɪnʲ əmʲɑːɹʲəʁkʰ	The weather will be beautiful tomorrow.
Sentence 8	Beidh an aimsear ag báistí amárach. WILL BE THE WEATHER RAINING TOMORROW. bʲɛʃ ənʲ aɪmʲʃɪɹʲ ɛʃ bʲɑːʃtʲiː əmʲɑːɹʲəʁkʰ	The weather will be raining tomorrow.

Using the discrete approach to emotion classification, the speaker was asked to produce each sentence in seven different discrete expressions (*neutral*, along with the affects *angry*, *bored*, *happy*, *sad*, *interested*, and *relaxed*), aiming for six repetitions of each expression. Out of an intended 336 tokens (8 sentences x 7 expressions x 6 repetitions), the final number of stimuli for the initial analysis came to 282 tokens. The subject was directed to record all repetitions with as little intonational variation between repetitions of the same

expression as possible but still allowed to record additional repetitions where he considered appropriate (e.g., sentence variants with more regionally common structures, and multiple versions of a given affect for a sentence). The sentences, word gloss and English translations are found in Table 1 above.

After recording, the data went through a preliminary review to scan for errors in the recording. The audio files most conducive to analysis were chosen and cut based on clarity of articulation (i.e., sentences with stutters were removed) and phonetic consistency (i.e., similarity in the realizations of phonemes). A few repetitions were excluded from this preliminary analysis due to variations in the pronunciation of words (e.g., *sin* ('that') being uttered as both [ʃɪnʲ] and [sənʲ]), but most such cases were additional stimuli added by the speaker and did not drastically reduce the number of tokens. These additional variants have been retained for potential future use.

This culled set of data was then reviewed for the analysis of data and parametrization. This review focused on identifying the recordings with intonational patterns that matched the version of the utterance produced by the ABAIR synthesizer (Ní Chasaide et al., 2017), which, at the time of writing, utilized FastPitch (Łańcucki, 2020) to extract f_0 from a training corpus, use those values to train a neural network, and generate the prosody contours at the synthesis stage. This training corpus of Irish data is what ensured that the ABAIR output matched the typical intonation patterns of the Kerry dialect (M. Dalton, 2008; M. Dalton & Ní Chasaide, 2007), as it also allows one to modify the pitch contour as desired based on the expertise of phoneticians involved in the project. This selection was done considering possible future applications of the present research, given that the practical implementation of parameter changes is more easily achieved if the data matches the natural intonational pattern of the dialect outputted by the ABAIR synthesizer. Through this review, a few issues were identified with the elicitation materials, and some of the recorded utterances were removed from the initial set (i.e., sentence sets presenting obvious focus shifts on most stimuli). Recordings of sentence 2 ('*Geobhmí bád ar an lár sin*') were selected as the most optimal for further experimentation given its intonational pattern and neutral semantic content. The accented syllables are '*bád*' and '*lár*', and the intonation contour of the recording was consistent with past description of high falling H*+L tones for this dialect (M. Dalton, 2008; M. Dalton & Ní Chasaide, 2007).

3.2. Data Annotation and Segmentation

The selected stimuli were imported to Praat (Boersma & Weenink, 2024) for segmentation and intonational annotation. The segmentation and annotation for the *neutral* and *angry* stimuli were conducted first and used as reference for other samples to ensure that the segmentations of all stimuli were matched in terms of phoneme and sentence boundaries (i.e., boundaries were marked roughly at the same pulse relative to neighboring phonemes, so that stimuli would be segmented as identically as possible depending on similarities between them). This was also done for the purpose of time normalization between affects down the line.

Preliminary intonational annotations were carried out using the IViE system for prosodic labelling (Grabe, 2001). While there is no prosodic labelling system specific to the Irish language, IViE was chosen due to it being based on the current standard for English intonation, ToBI (Silverman et al., 1992), but also allowing for comparison between multiple varieties of English, as previously mentioned. It has also been widely used for Irish prosodic annotation in previous studies (e.g., M. Dalton & Ní Chasaide, 2003; O'Reilly & Ní Chasaide, 2015; O'Reilly & Ní Chasaide, 2018). The annotations presented here, however, were kept much simpler than in such studies specifically focused on Irish intonation patterns. The IViE tone labels were used only at its canonical phonological tier to describe the utterance's intonational structure, as they were only conducted to simplify the pitch contours of these utterances into the IViE labels to facilitate the matching of intonation patterns between affects of the same utterance, as well as to the aforementioned typical dialect contours¹. It was decided that this would allow for identification of the best intonational pattern matches across affects, which would also facilitate the analysis and implementation of findings during the resynthesis phase. The intonational annotation of the *neutral* and *angry* affects served as a basis for the selection of sentence 2 as the most appropriate for the analysis, which showed matching patterns across both affects and all their available repetitions. The complete set of annotations up to this stage of the research can be found in Appendix A.

¹ Data review, intonation analysis and annotations were conducted with the guidance of Dr. Maria O'Reilly and Prof. Ailbhe Ní Chasaide, expert in-house phoneticians specialized in Irish intonation, to ensure a high standard of data selection for subsequent voice source analysis.

In addition to the IViE intonation annotations, other segmentation tiers were also added to the analysis on Praat (Boersma & Weenink, 2024) in order to limit each recorded sample to a single corresponding annotation/segmentation collection file. The phonetic annotation used on the phoneme segmentation tier was based on the IPA illustration of the Irish phoneme inventory (Ní Chasaide, 1999). At this point in the study, the segmentation was done in three different tiers as potential options for the subsequent time alignment of voice source parameter shifts applied globally and locally to add affective coloring to the neutral utterance at the implementation and resynthesis stage, given the differences in duration between the recorded neutral and affect-colored utterances. These tiers were: (1) chunks, i.e., dividing the utterance into its syllables; (2) metrical feet, i.e., combinations of two or more short or long syllables that form a basic unit of rhythm across the utterance; and (3) phonetic structure, i.e., breaking down the utterance into individual phonemes. For the global shifts, the start and end of the utterance were initially the most crucial time points. For the local shifts (aligned to the prosodic structure of the utterance and applied to the pitch accented syllables), on the other hand, the more detailed segmentations were important to test and determine potential anchor time points for these parameter shifts to be applied to, for which the phonetic tier was chosen as the most appropriate given how it provided precise time points for the start and end of the pitch accented vowel.

While the phonetic tier was ultimately used to determine anchor points for the local shifts as well as for the time-alignment of global shifts (explored in section 3.4 in more detail), the chunk tier was selected as the most appropriate for illustrating the alignment of voice source parameters to the intonational structure of the sentences (i.e., the changes in the voice source in relation to pitch accents and deaccentuated syllables of the utterance) more cleanly, and used for the creation of the graphical visualizations in section 4.1. These visualizations were created through a custom Praat script (Rodgers, 2022), further detailed in section 3.3.2, which also only allowed for a single annotation tier to be displayed on the horizontal axis, thus necessitating the choice. A segmentation example from Praat (Boersma & Weenink, 2024) can be seen in figure 7. In addition to the phoneme transcriptions, the capital 'B' seen in the figure refers to the burst sound of the plosives (a consonant produced by the interruption of the airflow at some point in the vocal tract filter, followed by a sudden

release of air), marked for the sake of detail and accurate delimitation of the following vowel, and technically part of the preceding consonant.

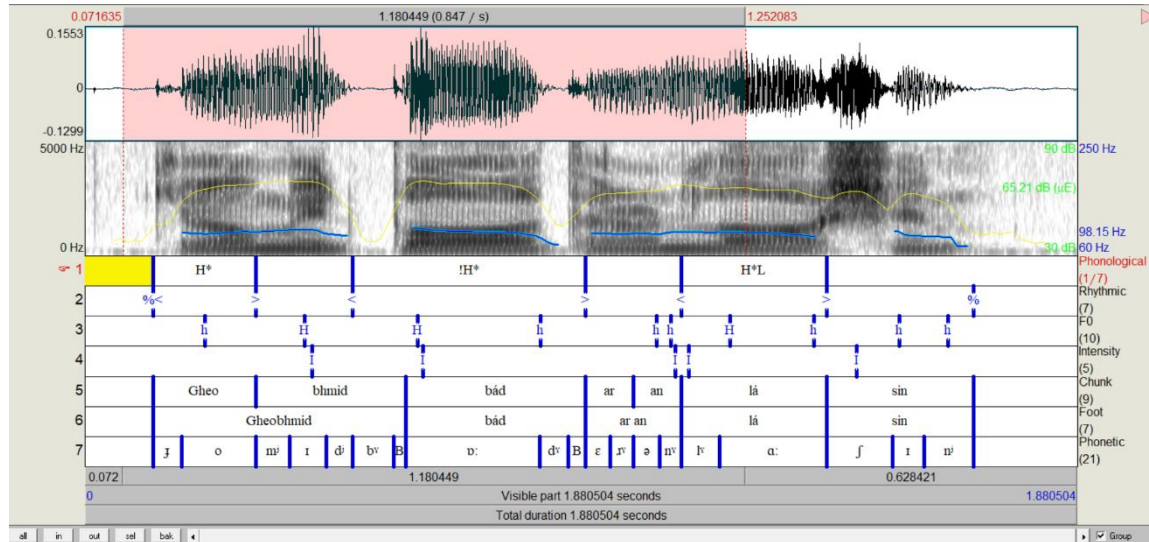


Figure 7 – Praat (Boersma & Weenink, 2024) Segmentation and lviE (Grabe, 2001) annotation of Sentence 2 *Neutral*, Repetition 1.

This more in-depth intonational analysis of data revealed one issue with the sentence 2 set: despite being the most consistent set in terms of matching intonation pattern to the neutral synthesizer output, a focus shift of the main pitch event from ‘*lá*’ to ‘*bád*’ was found in the *happy* affect stimuli. It was speculated that this was due to the data acquisition method of acted stimuli – the subject may have spoken the utterance picturing a specific scenario where the emotion of the speaker was directed towards one element in the sentence (‘*bád*’), rather than spoken in broad focus despite being advised not to do so.

This issue was identified much later in the research than ideal due to restrictions at the time of recording, including limited campus access and limited availability of the speaker. Such limited availability resulted in complications in scheduling new recording sessions in a timely manner. Additionally, it was considered that adding conscious constraints to the speaker’s dialectal prosody could divide his attention and affect his performance in terms of emotion portrayal. It was therefore decided to utilize the current stimuli for the study by applying a few measures to compensate for the focus shift in the *happy* stimuli of sentence

2, such as parametric analysis of other sentences and descriptive data from the literature to guide the parameter modifications at the resynthesis stage, along with experimental parametric adjustments to the production data, described in more detail in further sections.

3.3. Data Analysis and Visualization Methods

This section describes the methodology used for analyzing the recorded data through inverse filtering and parametrization necessary for the proposed experiments after segmentation and annotation were conducted, including initial approaches considered for this study. This analysis was conducted to create a guideline for the parameter shifts later applied to the neutral baseline, used to generate the stimuli presented to listener judges in a series of perception tests.

3.3.1. Initial Data Analysis Approaches and Considerations

The initial approach intended for data analysis was automatic voice source analysis through GlórCáil (Murphy et al., 2020) detailed in the review of the literature. However, upon conducting the voice source analysis and parametrization, a few issues were identified. The first related to the complicated nature of automatically extracting voice source parameters and accuracy of automatic voice source estimation and parameter extraction. Related errors were identified in the GlórCáil automatic analysis output, most notably an inversion of the RD parameter, as well as a capping of RK. While revisions have been done to the system, such issues led to the exploration of further alternatives.

One such alternative for the automatic analysis was the GloRí system (J. Dalton et al., 2014). Similar issues arose, along with odd fluctuations in the parameters. It was therefore decided to investigate stimuli more closely through a manual inverse filtering approach (Ní Chasaide et al., 1992), initially intended as a form of gold standard for the data analysis. This confirmed that the errors were not only an issue with the automatic analysis systems but also showed that this speaker's particular voice is very challenging to accurately analyze and parameterize. This was due to a natural roughness in his voice that resulted in a very jagged speech waveform, which made it very difficult to accurately mark the closed phase of each pulse, completely cancel formants in the spectrum, and match the resulting pulses to the LF model (Fant, 1995) pulse as closely as a more ideal voice would. A high degree of expertise with the manual source analysis system was therefore necessary to

compensate for this issue, and ensure the analysis was conducted as smoothly and as accurately as possible.

In terms of resynthesis, GlórCáil has successfully added emotional coloring to synthesized speech in previous experiments (Murphy et al., 2022). However, the stimuli from these studies were created by directly modifying speech through the system's code. The current interface is not equipped with time segmentation capabilities or direct value input and thus, unfortunately, it is not suitable for precise changes to parameter values at specified timepoints. Additionally, the resynthesized audio based on the automatic analysis also featured dramatically lower EE values than the original stimuli, degrading the quality of the output signal (something that is also mentioned by the author).

Given these issues, it was decided to ultimately utilize manual voice source analysis rather than any of the automatic methods available. Because this method is very time consuming, the manual analysis was conducted on a smaller set of data. This method, however, yields a much more detailed and accurate analysis of the voice source. The analysis is described in more detail in the subsection below.

3.3.2. Manual Analysis Methods

The manual analysis of data was conducted using the ISF System, an inverse filtering software described in detail in Ní Chasaide et al. (1992), Monahan (1996) and Gobl and Ní Chasaide (1999, 2010). While much slower than automatic approaches, the manual approach provides superior accuracy in voice source parameter estimation compared to automatic methods. For this, informal listening sessions were conducted, and a smaller number of stimuli was chosen from the initial batch through the same criteria of the previous selections: (1) the stimulus was the best at conveying the target affect within its batch; and (2) the stimulus matched the intonational pattern expected for the Kerry (Dingle) dialect without shifts in nuclear placement (i.e., high falling H*+L tones as described in M. Dalton (2008) and M. Dalton and Ní Chasaide (2007)). While most stimuli were acquired from the sentence 2 set (*'Geobhaimid bád ar an lá sin'*), one stimulus from sentence 4 (*'Beidh Barra ar an bád leo'*) was also analyzed for the *happy* affect because, as mentioned, all *happy* stimuli for sentence 2 presented focus shifts.

The chosen stimuli were down-sampled from their original 44.1 kHz to a 10 kHz sampling frequency (the required input for the ISF system) through Praat (Boersma & Weenink, 2024) and converted to .smp format via the Wavesurfer software (Sjölander & Beskow, 2000). The manual inverse filtering involves interactive fine-tuning of the frequencies and bandwidths of the anti-resonances used to cancel the effect of the vocal tract filter. This optimization of the anti-resonances is carried out for each glottal cycle to acquire an estimate of the differentiated glottal flow signal. Then, this signal is parameterized using a manual model matching technique, which involves fitting the LF model to the differentiated glottal flow pulses.

While the ISF system provides all the LF model parameters detailed in the review of the literature, the ones explored in more depth in the present research and therefore more thoroughly detailed below are f_0 (the fundamental frequency), EE (the excitation strength, which corresponds closely to the overall intensity of the source signal), and RD, the global waveshape parameter proposed by Fant (Fant, 1995, 1997). RD is not given directly through the ISF output but rather calculated subsequently.

After this analysis is concluded, the parameters are exported and combined with the Praat (Boersma & Weenink, 2024) segmentations described above. Segment durations of the non-neutral utterances were time-normalized to the segment durations of the neutral baseline using a custom Praat script (Rodgers, 2022). The script allows for visualization of parameters across multiple utterances through a normalization algorithm that allows for stimuli to be time-aligned according to the segmentation boundaries of a chosen stimulus (given that segmental durations and overall length differ between each recorded stimulus), facilitating visualization of such boundaries in relation to the neutral baseline.

With the manual analysis values and contours in hand, the visualizations presented in chapter 4 were used to help guide the decisions regarding the manipulations of parameter contours and segmentation points at which scaling factors were calculated for the resynthesis stage of this project.

3.4. Parameter Modification and Resynthesis Methods

Using the manual analysis of production data presented and literature findings as reference, a novel approach for modifying the parameters of the neutral baseline was created. This involved a method for separating the global and local shifts associated with each affect, and to apply such changes at both the global, local and combined (global + local) levels to the neutral baseline, presented here in general terms.

The global shifts were captured and applied through a linear regression analysis of the f_0 , EE and RD parameters. Initially, an overall, utterance-level regression line was acquired; then, a secondary linear regression analysis was carried out in a similar manner to the approach used in Haan, Van Heuven, Pacilly, and van Bezooijen (1997), using only the data points below (for f_0 and EE) or above (for RD) this utterance-level regression line. This secondary regression line is believed to better represent the global voice source characteristics, as it is less affected by the local contours associated with pitch accents. A visual example of this approach for EE can be seen in figure 8 below.

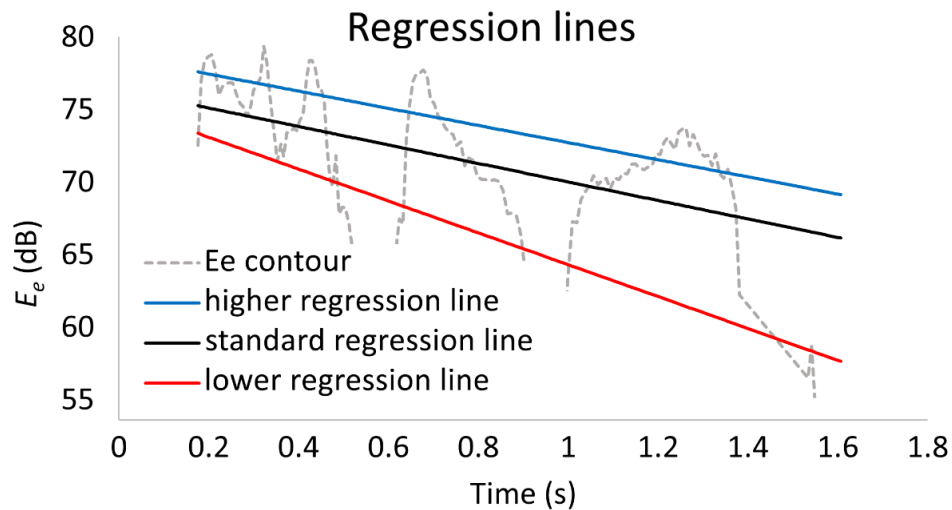


Figure 8 - Example of the secondary regression lines approach used for the global parameter manipulations, illustrating EE.

The local shifts were applied through an anchor-point system, developed through the visualization of parameter contours from the manual voice source analysis. These anchor points, based on the phonetic-level segmentations previously conducted in Praat (Boersma & Weenink, 2024), were used to extract measurements and calculate scaling factors between the neutral and the target affect utterances for the application of local shifts. Additionally, these anchor points were also used for time-alignment of the regression-based global manipulations, given that the neutral baseline utterance and the target affect utterances have different durations. Parameter contours between the anchor points were stylized through linear interpolation. A schematic representation of these anchor points and how they were used for time alignment is shown in figure 9.

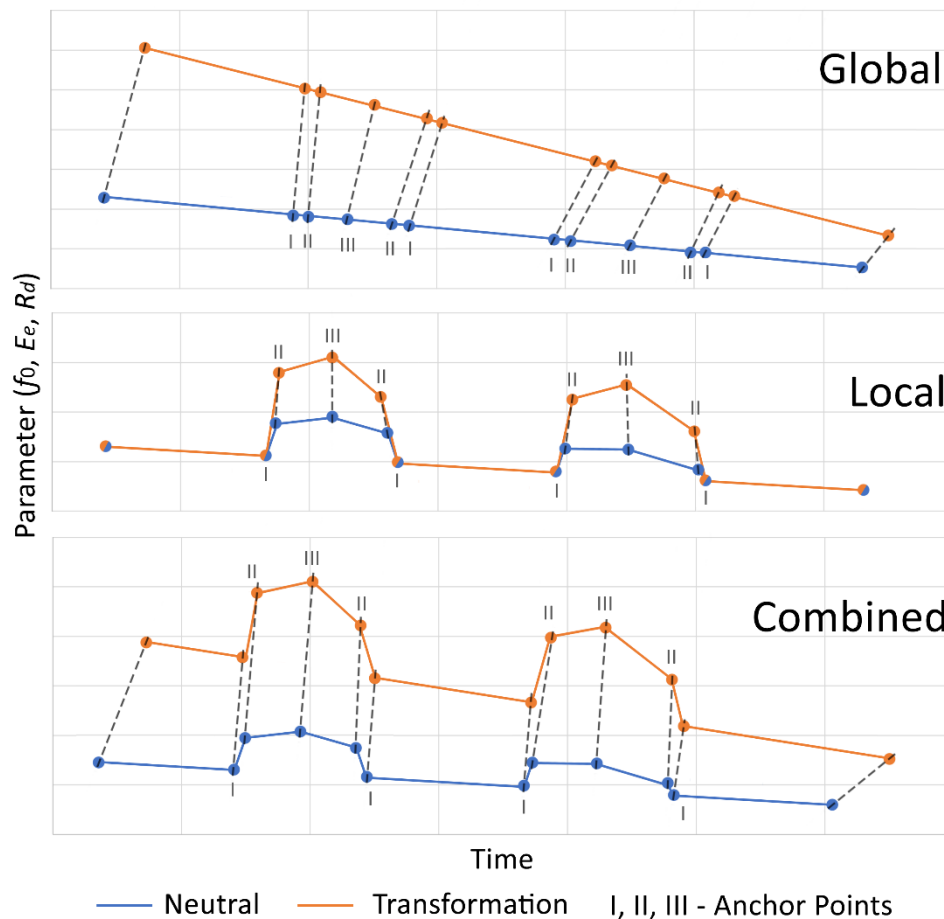


Figure 9 - Schematic representation of the voice transforms proposed. Anchor points are represented by the roman numerals I, II and III, used both for scaling of local shifts and for time-normalization of the global shifts.

Finally, the combined transforms bring together the global and local shifts, with the scaling factors of the local shifts being applied to the secondary regression line after the application of the global shifts. The results of the manual analysis, contour visualizations and more detail on the development of the modification system can be found in chapter 4.

To apply the desired affect-targeting manipulations and resynthesize the voice source parameters, the Voice Source Generator system (Z. Wang & Gobl, 2023a, 2023b), developed for the generation of realistic voice source signals, was used. It was made using the MATLAB App Designer (The Mathworks Inc., 2022), and its parametric voice transformation feature is realized using MATLAB's Live Script. The system takes the manual analysis parameters of the neutral baseline as input, and applies the desired manipulations based on the principles of the transformed LF model (Fant, 1995, 1997). Global shifts are done by inputting the regression coefficients and intercepts, and local shifts are done by inputting time points (which are also used for the time normalization) and target levels for each anchor point, which the system then uses to calculate the scaling factors.

The LF model, using the global waveshape parameter RD, was chosen for this research in both its analysis and synthesis stages, as it was found that the synthetic stimuli generated by it were perceptually closer to the target stimuli than other voice source models (Kreiman et al., 2015), which can help avoid distractions to listeners caused by less natural-sounding vocal stimuli. A more recent, aliasing-free version of the model is used in the VSG to account for aliasing distortion as described in Gobl (2021) and Z. Wang and Gobl (2023a). The model has also been extensively used in the literature and in the body of research that this project builds directly upon, as its small number of control parameters aligns with the goal of creating a simplified set of parameter changes associated with a limited number of affects. More detail on how the VSG applies the parameter manipulations based on the approach developed for the present research can be found in section 4.2.

3.5. Visual Stimuli Creation Methods

To explore the effects of the visual context on the recognition of emotion in the voice, a set of visual stimuli was created and used in Experiments 3 and 4 (chapter 7, sections 7.3 and 7.4, respectively). Rather than using common approaches in the literature such as hiring actors for shooting photographs and videos, these stimuli were created through 3D animation in order to create the expressions in a systematic and controlled manner than would allow for accurate time alignment to the neutral baseline and adjustments based on the tempo shifts of some vocal stimuli.

In terms of theoretical background, Paul Ekman Group's Facial Action Coding System (FACS) (Ekman, Friesen, & Hager, 2002) was used to establish the facial movements to be connected to each affect. The Action Units in the system quantify the muscle movements involved in the expression of emotion on the face and defines a neutral baseline for applying such modifications. However, because the FACS only explores the movement of individual muscles without directly connecting them to specific affect categories, a framework of facial expressions of emotion for sculptors by Zarins (2017) was used as an additional reference, allowing for the association of the Action Units to the target affects in the present research.

The software chosen to create and animate the facial expressions was Daz Studio (Daz Productions, 2022), a free 3D modelling and animation software that provides high-quality, pre-rigged realistic 3D human models and allows for manipulation of facial features through sliders, making such manipulations directly measurable. Likewise, lip-synching and eye movement animations, matched to the neutral utterance and added for the sake of naturalness, were also implemented via the animation tools in Daz Studio. Once finished, the animations were exported as individual frames and compiled in Adobe Photoshop (Adobe Inc., 2023) in video format to account for any necessary timing adjustments. Photoshop was also used to combine these visual stimuli to the vocal stimuli, creating the final stimuli combinations to be used in the perception experiments.

Before the combined vocal and visual stimuli were presented to listeners, the visual stimuli were pre-validated in a multiple-choice perception test to evaluate if the created animations successfully conjured the target affects for naïve viewers. The Action Units used

for each affect explored in the present research, their respective slider measurements in Daz Studio and the validation test for the visual stimuli are detailed in chapter 7.

3.6. Perception Tests

The vocal stimuli created for each of the experiments in this project were presented to listener judges in the form of perception tests, each conducted individually and with specific variations detailed in their respective sections.

Overall, tests were presented in an online format to Irish speakers with varying levels of proficiency. The online format was used to reach as many Irish speaking subjects as possible across longer distances, given the small pool of informants that met the research requirements. This format also allowed for a more flexible setup of the listening test, as well as the convenience of exporting responses directly to Microsoft Excel for data processing. The stimuli were uploaded and embedded to the tests in the form of YouTube videos. Listeners were asked to listen to the stimuli in a quiet room while wearing headphones, and to make a judgement of the emotion for each specific stimulus. The order of the stimuli was randomized for each test, and primer videos including all the stimuli in the test in sequence were presented after these instructions. Practice runs were also conducted before the test proper.

Responses were collected in two formats: (1) a discrimination-based approach with a subsequent strength rating for Experiment 1, and (2) a scale-based approach, presenting two opposing affects on each end of a bipolar scale with different strength scores for all subsequent tests, like the approach used in Gobl and Ní Chasaide (2003b).

3.6.1. Discrimination-based Task

The initial approach used in the first experiment consisted of an affect discrimination task. Stimuli were presented to listeners through Microsoft Forms in six sets, each set containing 16 distinct stimuli. The stimuli in each set were the same between each of them, but in different randomized orders. For each individual stimulus, listeners were asked to select the perceived emotion from the voice from a list that included all the synthesized affects (*angry, bored, happy, relaxed* and *sad*) as well as a *neutral/no affect* option. If an

affect was chosen, listeners were then taken to a follow-up question asking them to rate the strength of the emotion on a 5-point scale from weak (1) to strong (5).

This discrimination approach was based on a previous study by Murphy et al. (2022), who also investigated the role of global, local and combined voice source shifts in the expression of affect by using a multiple choice task with subsequent intensity ratings in Microsoft Forms. This format was initially selected for Experiment 1 because the multiple-choice task was considered less tiresome for participants (given the total number of questions presented to them), and because the separation of affect identification judgements from the intensity scores was initially deemed easier to interpret for researchers.

Questions and answer options were presented in Irish, and the responses to the first set of stimuli were discarded from the analysis of results and considered the practice run for this task.

3.6.2. Scale-based Task

A scale-based test, an approach previously used in the study of attitude (Uldall, 1964) and affect (Gobl & Ní Chasaide, 2003b), was used for experiments 2-4, as it allows one to measure both the activation of the expression and intensity at the same time. The reasoning for this methodological shift were results from Experiment 1 (detailed in chapter 5), which showed a tendency of most listeners to select mid-range strength scores consistently after selecting an affect from the multiple-choice options, and considerations about discrimination tasks detailed in section 2.3.3. Additionally, the method also allowed for the investigation of whether the presented stimuli conjured an affect other than its target if listeners were presented with response options that did not include the stimulus' target affect.

These tests were presented to listeners through Google Forms, chosen for its increased stability and flexibility. The same set of instructions was given in both Irish and English. Primer videos were presented, consisting of the neutral stimulus presented individually, followed by the set of all stimuli to be presented in the test. They were asked to use these primers as an opportunity to adjust the volumes of their headsets.

The scale-based tests were divided into three tasks, each with a different scale of affects. A practice run was presented for the first task with varying number of stimuli depending on the experiment, for listeners to familiarize themselves with the test format. Each task presented them with all the affect-targeting stimuli in the set, always preceded by the neutral stimulus as a pair of utterances per video/question.

The tasks asked listeners to judge the emotion of the second utterance of the pair on a 7-point scale of intensity between -3 (low activation) and +3 (high activation), each polarity representing one of two opposing affects: *sad* vs. *happy*, *bored* vs. *interested*, and *relaxed* vs. *angry*. The ± 2 represented a moderate expression of the affect, and ± 1 represented a slight expression. Listeners were asked to choose the midpoint 0 (zero) whenever no affect was perceived (no emotion/neutral), or when they perceived an affect from the stimulus that was not at any end of the scale presented for that subtest. Although there were no vocal stimuli targeting *interested*, the *interested* option was included in the scales as a counterpart to *bored*. In contrast to the discrimination-based task, because this test format already included a dedicated section for listeners to practice the task (and responses to these practice runs were not included in the statistical analysis), no set of responses outside of the practice run section was discarded.

3.7. Statistical Analyses

The test results were extracted from Microsoft and Google Forms as Excel sheets and cleaned through a custom MATLAB script² to remove any unwanted text from the direct test results output (which included listener instructions, section titles and questions), leaving only the relevant information for the statistical analyses (i.e., variable names such as target affect, perceived affect, and numerical intensity scores). The same script also incorporated an algorithm to turn the clean data into some of the figures used to illustrate affect recognition and perceived strength ratings (shown and credited in the respective experiment chapters).

Statistical analyses were conducted using R (R Core Team, 2021) through the R Studio environment (Posit Team, 2025), with the assistance of an expert statistician³. Each

² Script written by in-house researcher Zihan Wang, creator of the VSG system (Z. Wang & Gobl, 2023b).

³ Statistics consultation provided by Dr. Daniel Dempsey, PhD in Statistics from Trinity College Dublin.

task required different model specifications ensure a robust analysis that supported the perception test data. Methods are presented here in general terms, and specifics about the R packages and models used for each experiment are detailed in the respective experimental sections.

For the discrimination task, a mixed-effects generalized linear logistic regression model was used for the affect identification analysis (i.e., exploring which target affect and voice modification types were correlated with the perception of 'correct' target affect recognition), done on an overall basis across the two variables (target affect and voice shift type) and on a per-emotion basis. The dependent variable (perceived affect) was coded as a binary outcome due to issues with rank deficiency and data sparsity, which led to this choice of methodology. Factor variables were relevelled with 'neutral' (target affect) and 'none' (voice shift type) as reference variables.

Intensity ratings for this task were analyzed using a mixed-effects ordinal logistic regression model, individually examining the effect of target affect categories and voice shift types on listeners' intensity ratings, and the interaction between these two variables. Finally, a smaller mixed-effects ordinal logistic regression analysis was conducted to see if 'correct' affect identification (target affect = perceived affect) correlated with higher intensity ratings. For all models used in this task, participant ID was included as a random intercept. To control for multiple comparisons, post-hoc Holm-Bonferroni corrections (Holm, 1979) were also used, and p values below .05 post-correction were considered significant across the analyses.

For the scale-based tasks, mixed-effects ordinal logistic regression models were used, and model fitting was done on a per-scale basis. The models used the 7-point scales as ordinal dependent variables, analyzing the effects of the independent variables (target affect and voice shift type or visual stimulus combination type, depending on the respective experiments) and their interactions on affect identification and intensity ratings. Random intercepts for listener ID were included to account for individual differences, and Holm-Bonferroni corrections were also used to account for multiple comparisons. Confidence intervals were calculated using standard errors from the model output and p values below .05 post-Holm-Bonferroni adjustments were considered statistically significant.

Due to the format of the experiment, an overall model across all scales could not be fitted due to numerical issues and quasi-complete separation of data. The per-scale model was deemed the most robust and appropriate for the data after extensive testing, and the analysis was conducted so that significant values reflected the effectiveness of the independent variables in both affect recognition and intensity scores.

For experiment 4, small adjustments were done to the models to account for additional incongruent conditions presented as vocal and visual affect combination types, but models and analysis formats were the same as experiment 3.

4. THE AFFECT MODIFICATION SYSTEM

This chapter presents the results of the production data analysis carried out through the ISF system, along with contour illustrations used to inform the creation of the affect manipulation system. It also includes the detailed methodology for creating the affect modification system through the VSG system (Z. Wang & Gobl, 2023b), exploring how it applies the desired voice source shifts to the neutral baseline.

4.1. Production Data Analysis

This subsection presents the results of the manual voice source analysis in the form of contour illustrations⁴ for each of the relevant voice source parameters. These analyses were the basis for the modification system detailed in section 4.2., presented in further detail in subsequent chapters related to each of the experiments in the present research. The jaggedness of parameter contours seen below are a result of the aforementioned challenges in manually inverse filtering this speaker's particular voice.

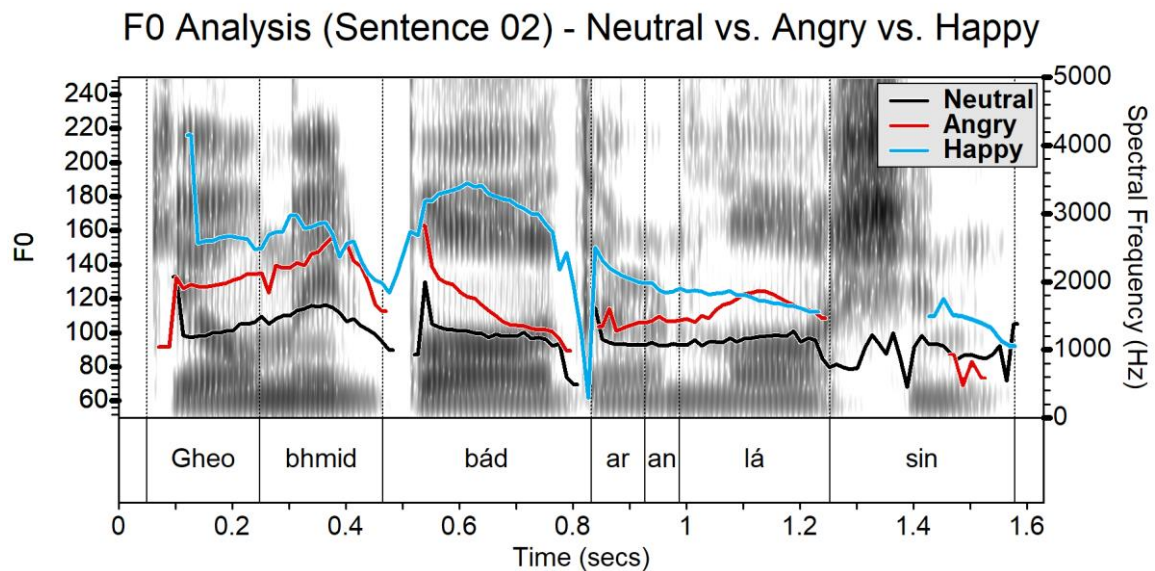


Figure 10 – F_0 contours from the manual voice source analysis comparing the *neutral*, *angry* and *happy* affects.

⁴ Parameter illustrations shown in this chapter were created through a custom Praat script by Rodgers (2022) and edited through Adobe Photoshop (Adobe Inc., 2023) for additional formatting adjustments.

F_0 contours acquired from the manual analysis of the recorded data for the neutral and target affects are presented in figures 10 and 11. Immediately noticeable in figure 10 are the distinct contours for the high activation affects *angry* and *happy*. *Angry*, as expected, shows a high falling peak on the accented vowels, especially on *bád*, where the fall is very steep. Globally speaking, *angry* f_0 levels are also higher than the neutral. *Happy*, on the other hand, shows a very high, rounded f_0 contour on *bád*, which is very different from the *angry* stimulus. F_0 in *happy* is also globally higher than in *angry*. The second pitch accent does not show any differentiation, the contour being instead a continuous fall from the start of the voiced segment ‘*ar an lá*’, since for this utterance all stimuli of the *happy* affect presented focus shifts. This issue was addressed by manually analyzing an additional stimulus of another sentence.

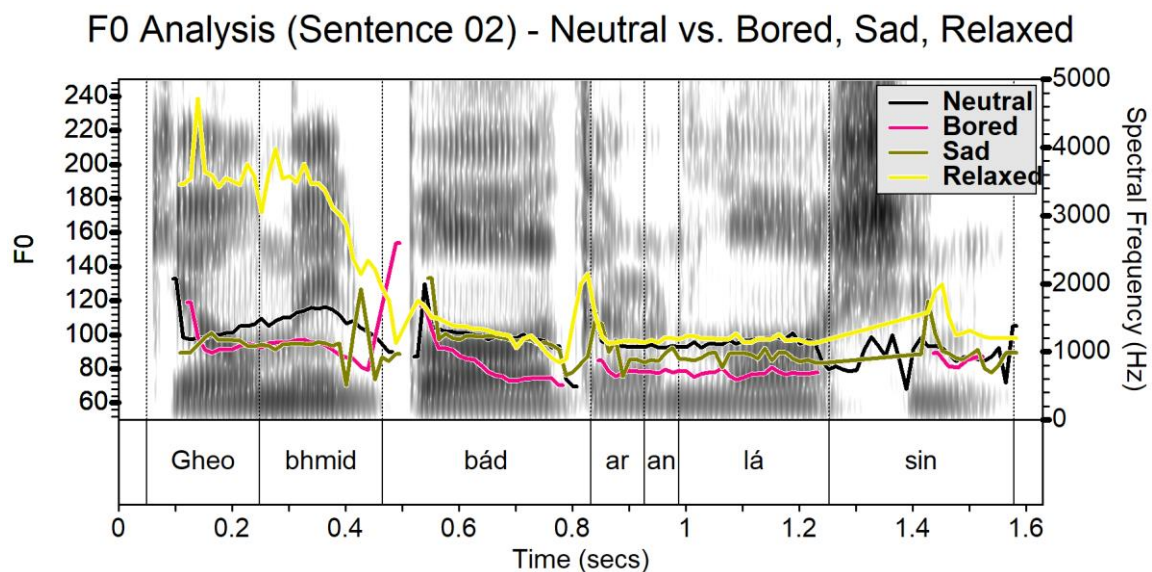


Figure 11 - F_0 contours from the manual voice source analysis comparing the *neutral*, *bored*, *sad* and *relaxed* affects.

For the low activation affects (figure 11), differentiation from neutral was much more subtle. As mentioned, expert listeners involved in analyzing the intonation patterns of these sentences perceived the neutral produced by this speaker as more on the low activation side and closer to what one would consider a ‘*sad* voice’ than a neutral synthesizer output. This partly explain the closeness of the contours in figure 11, along with

the fact that these affects, according to the literature, tend to have low parameter variability. It was, however, decided not to arbitrarily modify the neutral as produced by the actor, and to keep the baseline as the original recording.

Bored shows the lowest overall f_0 level, with a flat contour except for the steeper fall on the first accented syllable. *Sad* maintains a very flat contour, also lower than neutral but for the accented syllable *bád*, where it matches neutral closely. Finally, *relaxed* presents very high f_0 at the start of the utterance, only to sharply fall back into a contour that is also very close to neutral from the first pitch accent onward. For an affect considered a low activation version of *happy*, this significant difference from the *happy* contour was quite surprising, as the recording was quite successful at portraying the affect.

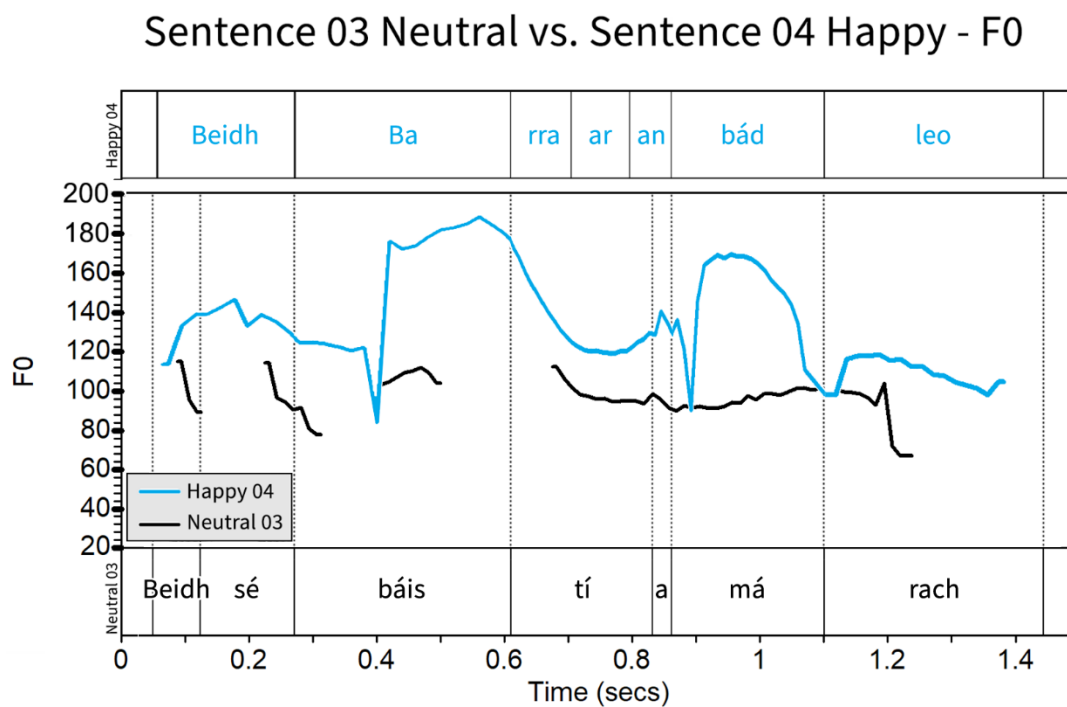


Figure 12 - F_0 contour from the manual voice source analysis comparing sentence 3 *neutral* ('Beidh sé báistí amárach') and sentence 4 *happy* ('Beidh Barra ar an bád leo'), time-aligning the pitch accents between utterances.

Additionally, a comparison between the *neutral* stimulus of sentence 3 and the *happy* stimulus of sentence 4 was also created (figure 12) due to the focus shift presented

in all *happy* stimuli of sentence 2, causing heavy deaccentuation of the second pitch accent (*lá*). To acquire what was considered to be the proper contour for the *happy* affect in the accented syllables, additional sentences were manually analyzed.

Sentence 3 was used for visualization purposes only due to the similarities to sentence 4 in terms of length and locations of pitch accents in relation to the non-focal material. When the second nuclear accent follows the dialect's intonational patterns, the same curved peak for the *happy* affect seen on the first nuclear accent of sentence 2 is present. Therefore, it was decided to transplant this contour onto the second nuclear accent of sentence 2 for the manipulation and resynthesis stage. For the calculation of scaling factors described in the next section, only the second nuclear accent of sentence 4 was used on top of sentence 2 for all modifiable parameters.

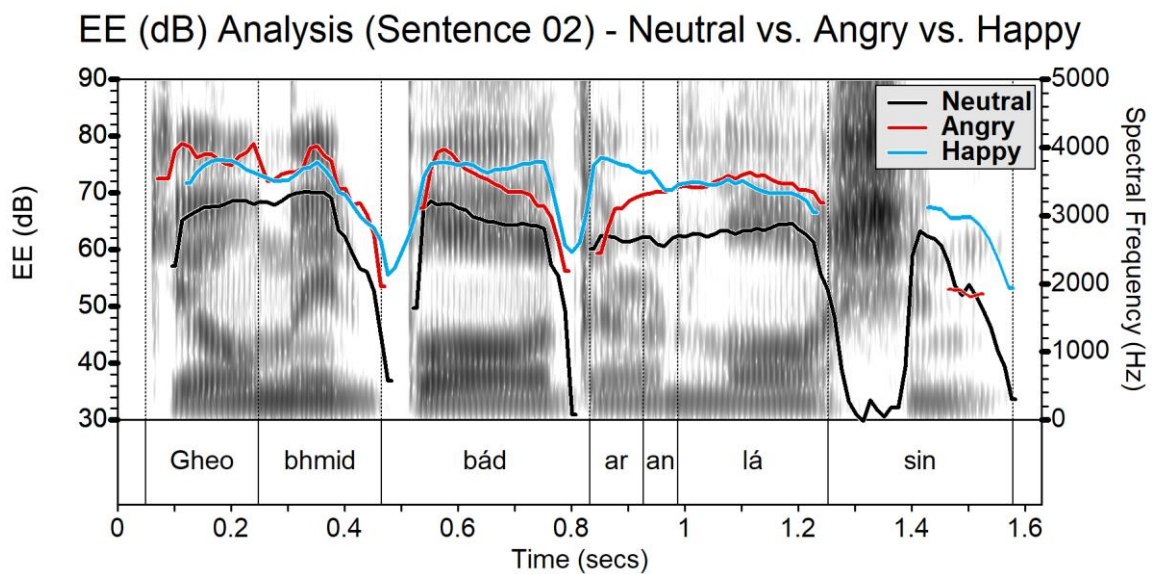


Figure 13 – EE contour from the manual voice source analysis comparing the *neutral*, *angry* and *happy* affects.

For EE, the high activation affects (figure 13) show some similarities to f_0 . *Angry* has a similar high fall on *bád*, also seen on the *lá* vowel to a lesser extent. *Happy* does not feature the same rounded contour as f_0 but still maintains a steady high contour throughout the vowel in *bád*. Again, the second pitch accent is not representative of the target contour. As expected, both affects show higher global EE levels than neutral.

EE (dB) Analysis (Sentence 02) - Neutral vs. Bored, Sad, Relaxed

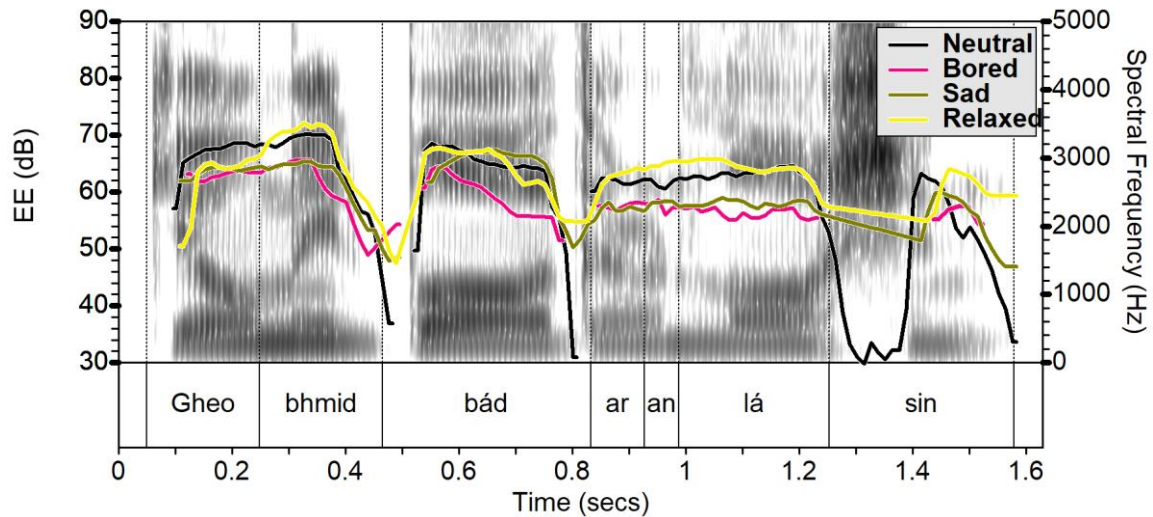


Figure 14 - EE contour from the manual voice source analysis comparing the *neutral*, *bored*, *sad* and *relaxed* affects.

Likewise, *bored* and *sad* EE contours (figure 14) are similar in shape to f_0 and close to the neutral. *Bored* again has the lowest overall levels, with a sharp drop on the first accent and a steeper slope towards the end of the utterance, while *sad* is flatter than neutral, matching the latter on the first accent but also with a steeper declination slope. EE for *relaxed* does not have the high utterance-initial rise like f_0 , and levels stay close to the neutral for the most part, with only a slight rise on the nasal and lateral (*an lá*).

Figure 15 below shows the EE contour (in dB) comparisons between sentence 3 and sentence 4 for the *happy* affect. These extra stimuli further confirm the continuous EE rise on the pitch accents, again like f_0 , but with a flatter contour. Because of this, it was considered appropriate to transplant the contour seen in '*bád*' for sentence 2 onto '*lá*', to compensate for the focus shift in the recordings.

Sentence 03 Neutral vs. Sentence 04 Happy - EE (dB)

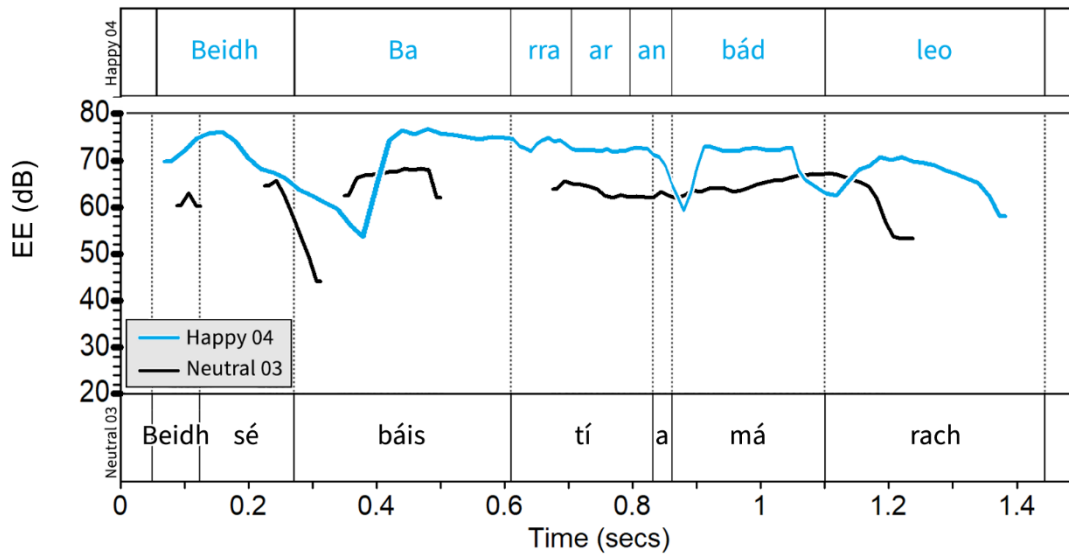


Figure 15 – EE (in dB) contour from the manual voice source analysis comparing sentence 3 *neutral* ('Beidh sé báistí amárach') and sentence 4 *happy* ('Beidh Barra ar an bád leo'), time-aligning the pitch accents between utterances.

RD Analysis (Sentence 02) - Neutral vs. Angry vs. Happy

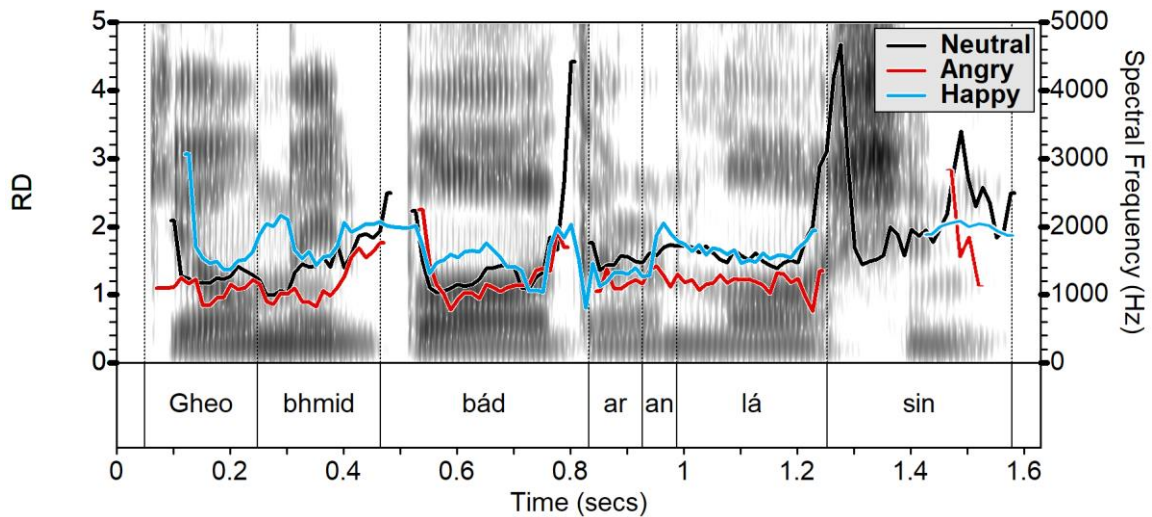


Figure 16 - RD contour from the manual voice source analysis comparing the *neutral*, *angry* and *happy* affects.

Changes in RD between affects are more subtle than the other two parameters for the high activation affects (figure 16). As seen, *angry* shows an overall tenser voice quality (lower RD), without large contour differences from the neutral other than a slightly more accentuated rise on *bád*. *Happy*, on the other hand, presents a laxer voice quality at the start of the utterance, later falling to match the neutral. Accented syllables are only slightly tenser than the non-focal material, with a tensing on ‘*ar an*’ in particular.

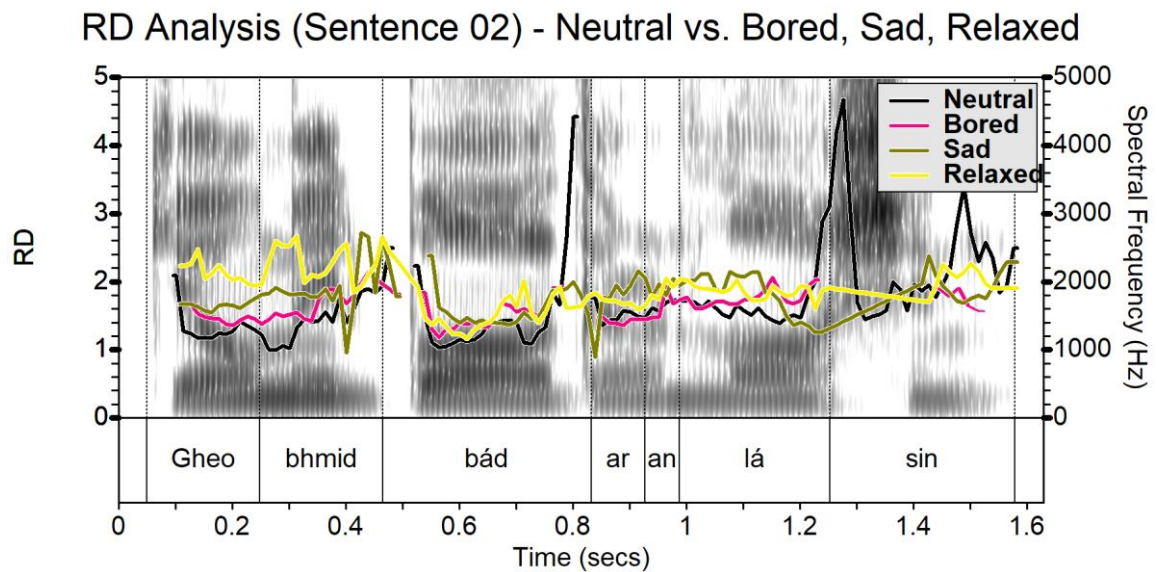


Figure 17 - RD contour from the manual voice source analysis comparing the *neutral*, *bored*, *sad* and *relaxed* affects.

Accordingly, low activation RD contours (figure 17) show small differentiations from the neutral. *Bored* is slightly laxer than neutral, especially in the pitch accents. *Sad* is slightly higher/laxer than *bored*, returning to the neutral on *bád*. Finally, *relaxed* starts with a very lax voice quality (high RD), drops closer to *neutral* on *bád*, and resumes with a laxer but flatter contour.

Lastly, the extra *happy* stimulus analysis shows a tenser voice quality on both pitch accents for sentence 4 *happy* (figure 18), which, therefore, will follow the same methodology as the other parameters in the resynthesis phase.

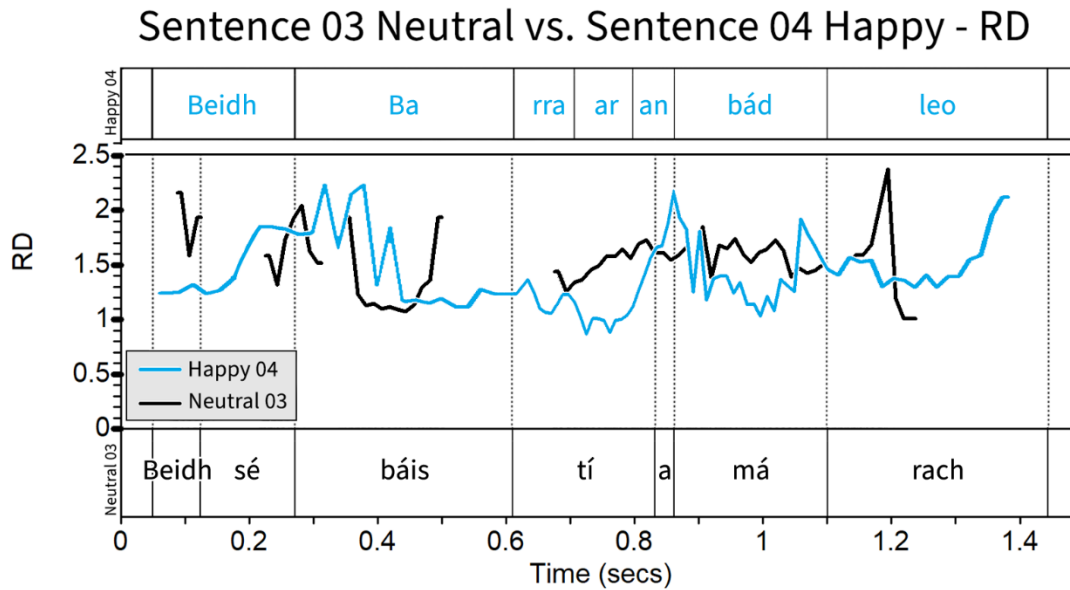


Figure 18 - RD contour from the manual voice source analysis comparing the sentence 03 *neutral* ('Beidh sé báistí amárach') and the sentence 04 *happy* ('Beidh Barra ar an bád leo'), time-aligning the pitch accents between utterances.

The *interested* affect was excluded as there were no versions of the *interested* stimuli across all recordings that did not contain some degree of focus shift. It is hypothesized that this is because this emotion is highly context-dependent, and will inherently contain focus (i.e., a speaker is interested in a certain element of an utterance, not in the content of the utterance as a whole). It was therefore decided to leave this affect out of the experiments, as the inherent focus shift added another layer of complexity to the overall aim of the project.

As mentioned, these parameter track visualizations from the production data were the main references to guide decisions regarding the stylization of the parameter contours and segmentation points for the calculation of scaling factors at the resynthesis stage. Additionally, features from the general literature on the topic of voice source correlates of emotion were also used to inform additional manipulations.

The case of the Happy stimulus

The *happy* stimulus was particularly challenging due to the focus shift realized by the speaker in all *happy* stimuli of sentence 2 (in contrast to the other affects, which

presented the typical Kerry intonation pattern), which removed focus from the second nuclear accent 'lá'. It was thus decided that the rounded f_0 contour, as well as EE and RD contours seen on the first nuclear accent of sentence 2, would be transplanted based on the manual analysis conducted on an alternative *happy* sentence with the proper intonation pattern (sentence 4).

To test the hypothesis that this transplanting would be successful, a preliminary resynthesis of the neutral baseline manipulated through the Voice Source Generator (Z. Wang & Gobl, 2023b) into the *happy* affect was done based purely on the values of sentence 2's *happy* stimuli production data. Then, a manual manipulation of the f_0 contour on the second nuclear accent was conducted on Praat (Boersma & Weenink, 2024). Informal listening sessions with Irish speaking colleagues confirmed that, even though only f_0 was altered (EE and RD were kept at production data values), the new *happy* stimulus was much closer to what would be expected of the Kerry dialect, in addition to sounding closer to the target affect. It was therefore decided to take these additional measures with the *happy* stimuli to control for broad focus in the stimuli presented to listener judges.

In the end, as described more thoroughly in the resynthesis methodology below, this transplanting process between the two pitch accents was extrapolated to the other affects as well for the sake of simplifying the sets of local changes required to achieve each of the affects. Global manipulations for *happy* were also based on sentence 4, since the narrow focus of the sentence 2 *happy* stimulus could affect values. A thorough description of the final methodology for manipulations is presented below.

4.2. Parameter Modification and Resynthesis Methods

4.2.1. The Parameter Modification System

Once the production data analysis was concluded, a method for modifying the neutral baseline was developed. The broad approach was based on the work carried out by Murphy et al. (2022) as well as production data from previous experiments (Gobl et al., 2002; Gobl & Ní Chasaide, 2003b; Murphy et al., 2017; Yanushevskaya et al., 2009, 2018) and previous descriptions of vocal parameters of emotion in the literature (Banse & Scherer, 1996; Ladd et al., 1985; Scherer, 2003, 2021; Scherer et al., 1984). At first, global shifts were done by modifying the mean, standard deviation, and z-scores (range) of parameters calculated

from the manual analysis described previously. However, such global measures may not be the best representations of global utterance trends, especially considering the effect of outlier pulses near stops and fricatives. Median values for f_0 were also explored as a potential method of excluding outlier pulses, but the question of how to effectively classify which pulses were outliers remained.

Eventually, declination shifts were considered (i.e., declination trend lines) as a better representative of a halfway trend between peaks and valleys. Linear regression analysis was considered an appropriate representative of a mean declination line. Shifts were created using the coefficients and intercepts of the regression analysis, along with a time normalization algorithm to properly align the trendline between the *neutral* and the affective stimuli given their durational differences. Furthermore, rather than using the main regression line for the global shifts, a lower regression line acquired by running a secondary regression analysis on the data points below the main regression line was used, similar to the approach used by Haan et al. (1997). This lower regression line was hypothesized to better reflect global parameter trends than the standard regression line, as the values of the former were not as affected as the latter by the larger parameter shifts seen in the pitch accented syllables (which are considered local shifts).

The local manipulations relied on a set of anchor points, which needed to be refined to avoid data points too close to the unvoiced areas (e.g., glottal stops) bordering the pitch accented vowels and that deviated too much from the vowel parameters. The intention was to choose only a handful of points within the vowel and stylize the contours between these points – i.e., the parameters would be interpolated into straight lines between these anchor points, to simplify the contour and therefore the manipulations for the future implementation of findings. The points chosen as the anchors were: (I) the first and last voiced pulses of the vowel in the accented syllables (*bád* and *lá*); (II) pulses located 20% into the vowel from the anchor points described in (I); and (III) the middle point of each accented vowel, i.e., the pulse closest to the midpoint between the anchor points described in (I). Other anchor positions were also considered based on the contours seen in the manual analyses, such as pulses 1/3 and 2/3 into the vowel depending on the target affect; in the end, however, it was decided to simplify these manipulations and keep to the same anchor points across all affects.

Finally, it was decided to define the local shifts as the deviations of the anchor pulses from its respective global trendline – i.e., the local shifts were deviations of the target pulse values from the target secondary regression line. Anchor points described in (I) were used to ‘anchor’ the parameters back to the parts of the utterance where the application of a local shift was not desired, whereas anchor points (II) and (III) were scaled accordingly. These relative shifts were considered the best definition of local parameter changes without being affected by the global shifts. In the end, the chosen approach was (1) linear regression analysis for the global manipulations; and (2) scaling factors at anchor points relative to the neutral and target affects regression analysis, detailed in following subsections.

Additionally, there was the issue of differences between pitch accents. Such differences were not as dramatic as expected from the regression lines, but, for some parameters, the second accent was weaker due to declination and, therefore, scaling factors calculated from the recorded data for the second pitch accents were much smaller than the first. For the sake of simplification as per the aim of this research, a single set of local scaling factors was created for each affect that would be applied to both pitch accents (like the transplanting approach mentioned previously). This involved finding a middle point between two respective scaling factors and manually adjusting the changes through many iterations of stimuli and careful listening and visualization to capture the emotion of the original recording as accurately as possible.

4.2.2. Application of Parameter Shifts and Resynthesis through the VSG System

As mentioned, the system used for the resynthesis and modification of voice source parameters was an aliasing-free version of the Voice Source Generator (VSG) (Z. Wang & Gobl, 2023a, 2023b), which applies parameter manipulations to a neutral baseline based on the principles of the transformed LF model (Fant, 1995, 1997) in the frequency domain (Gobl, 2021). The VSG interface is shown in figures 19 and 20 below.

VSG - voice source transformation - Downdrift 2 - additive

Introduction

This live script is to implement voice source transformation (on Ee, Rd and F0) on the resynthesised speech of the IFS analysis output, as a feature of the VS generator.

Table of Contents

- Introduction
- 1. Import and inverse
- 2. Resynthesis the source
- 3. Voice source transformation
 - 4.1 Resynthesis the original speech
 - 4.2 Resynthesis the transformed speech

1. Import and inverse

Import voice source parameters and formants.

Do inverse filtering and get the T.

filename = ;

Taking F0-Ee-Rd algorithm or F0-Ee-Rd-OQ algorithm?

paramAlgorithm = ;

Click Import and inverse.

2. Resynthesis the source

Generate the LF source signal, obtain the residual and calculate the scaling parameters of the residual.

Which version of LF model will be used here?

LFVer = ;

Press the button when you are ready to generate the source signal.

Optional: click if you want to play and plot the generated LF signal.

3. Voice source transformation

Global shift, local shift, or the combination of both?

shiftApproach = ;

For global shift: SKIP if using only local shift.

- Now taking the down-drift approach.

Please specify the first and last GCI (LF_Te in ISF output) of the target utterance.

Te1_T = ; % ENTER the first GCI of the target affect

TeT_T = ; % ENTER the last GCI of the target affect

Te1_0 = params.GCI(1); TeT_0 = 1.202; %time point for ABAIR sample; change for other baselines

Now specify the segment timepoints.

timepoint_0 = ; % Syllabic-level timepoints of the original utterance

timepoint_T = ; % Syllabic-level timepoints of the target utterance

Figure 19 - Part 1 of the VSG Interface from introduction, importing parameters and resynthesizing the neutral baseline, to the first part of the voice source transformation.

Now specify the coefficients and intercepts of F0, Ee and Rd of the original utterance and the target utterance.

xxxx_O: data for the original utterance (for now, neutral); xxxx_T: data for the target utterance

F0_coefficient_O = 9.9526 ;
F0_intercept_O = 91.5768 ;
F0_coefficient_T = -28.6087 ;
F0_intercept_T = 100.2468 ;
% NOTE: Ee in dB scale %
Ee_coefficient_O = 0.3714 ;
Ee_intercept_O = 62.9784 ;
Ee_coefficient_T = -4.049 ;
Ee_intercept_T = 59.4458 ;
Rd_coefficient_O = -0.6722 ;
Rd_intercept_O = 2.3357 ;
Rd_coefficient_T = 0.0844 ;
Rd_intercept_T = 1.6623 ;

For local shift: SKIP if using only global shift.

Please indicate the **scaling factors** for F0, Ee and Rd.

- For each parameter, please indicate the scaling timepoints and the factors.
- Please also ensure that the number of timepoints and the number of factors are the same.
- separate the timepoint values by **semicolon** or **space**.

F0_scale_TimePoint = 0.5431 0.5616 0.6373 0.7 ;
F0_scale_TimePoint = str2num(F0_scale_TimePoint);
F0_val = 0.91 0.50645 88.635435 7 ;
F0_val = str2num(F0_val);
Ee_scale_TimePoint = 0.5431 0.5616 0.6373 0.7 ;
Ee_scale_TimePoint = str2num(Ee_scale_TimePoint);
Ee_val = 0 1128.81405 869.83212 6 ; % NOTE: the measurements are in linear scale.
Ee_val = str2num(Ee_val);
Rd_scale_TimePoint = 0.5431 0.5616 0.6373 0.7 ;
Rd_scale_TimePoint = str2num(Rd_scale_TimePoint);
Rd_val = 0 1.36176 1.5381 1.68040 ;
Rd_val = str2num(Rd_val);

Press the button when you are ready to transform the source signal.

Transform the source

Optional: click if you want to play and plot the transformed voice LF source signal.

Play and plot the transformed voice source signal

4. Formant synthesise

4.1 Resynthesis the original speech

Play and plot the resynthesised signal

origName = ABNEUT_resynthtest_180 ;
Export transformed signal

4.2 Resynthesis the transformed speech

NOTE: The absolute value of the resynthesised signal amplitude is capped at 2^{15} , i.e. 32768 :')

Play and plot the transformed resynthesised signal

Export the transformed speech signal.

transName = ABbored_combined_1207 ;
Export transformed signal

Figure 20 - Part 2 of the VSG Interface, continuing from the voice source transformation parameter input to formant and final resynthesis.

To potentially achieve better sound quality, the glottal flow residual is modelled following the method described in Z. Wang and Gobl (2022), and is included as part of the source signal. However, in this research, in order to avoid unwanted artefacts produced by excessive residual, only the residual signal in the unvoiced regions (automatically identified as the pulses where EE is smaller than 1000, or 30 dB) was added to the LF source signal. This maintains the quality of the components not captured by the LF model (e.g., fricatives and stops, produced apart from the voice source), further elaborated upon in section 4.2.7.

In principle, to run the transformation, the user is required to specify scaling factors for each of the voice source parameters f_0 , EE and RD. The new parameters for the transformed stimulus are obtained using these scaling factors, and a new source signal is generated. The transformations carried out in the VSG system for this project are of three kinds: global (applied across the entire utterance), local (applied only to specific points of the utterance) and combined (global and local transformations added together). A more detailed description of the practical implementation of the transformations is presented below. Images illustrating the three different types of manipulations (global, local, and combined) are also presented, representing the first set of stimuli considered successful enough for an experiment to be conducted with listener judges. They utilize the recorded neutral as the baseline for the manipulations and calculations of the scaling factors, and to illustrate differences between the baseline and the transformed signal.

4.2.3. Initial attempts and considerations

Before generating stimuli for the experiments, the accuracy of the system was tested. First, images were created comparing the parameters of the original recordings and the output of the system for each affect using the same Praat script described during the analysis stage (Rodgers, 2022). Figures showed that the output of the system was satisfactory: apart from small fluctuations caused by the PSOLA algorithm (Moulines & Laroche, 1995), which is a normal occurrence in the process, local shifts were reaching the desired parameter levels. Global values were also sensible based on plots and numerical values.

Mean Absolute Difference scores were also acquired to numerically measure the accuracy of the output on the accented syllables in comparison to the original recordings (table 2). This was done for *angry* and *happy*, which were the two affects with the most

dramatic shifts from the baseline both in global and local terms, using f_0 and intensity parameters extracted from Praat (Boersma & Weenink, 2024) as a separate system from the VSG for testing how its own algorithm would read the VSG output. Because there were differences in the quantity of data points between resynthesized stimuli and recordings, data interpolation was conducted through MATLAB for the MAD analysis to be conducted.

Table 2 - Mean Absolute Difference values of each accented vowel for the *angry* and *happy* affects, for f_0 (in Hz) and intensity in (dB) extracted from Praat (Boersma & Weenink, 2024), testing the VSG system’s accuracy at applying the desired local shifts.

Affect	F_0 (Hz)		Intensity (dB)	
	Accent 1 (bád)	Accent 2 (lá)	Accent 1 (bád)	Accent 2 (lá)
Angry	2.1	18.1	2.6	4.3
Happy	6.1	15.4	2.0	12.1

Table 2 shows the intended applications of the local shifts; f_0 differences amount to only a few hertz on the first pitch accented vowel, which can be explained by the stylization of the contour as well as natural fluctuations caused by PSOLA. Intensity differences are slightly bigger, but still acceptable given the stylization.

As expected, due to the compromise between the two accents in terms of scaling factors (but with the shape of the first accent still being rather dominant, as seen especially for intensity values in figure 24), the changes to the second pitch accent are more prominent for both parameters. This also considers the existence of slight differences on the non-focal material due to the application method of global shifts, as the stimuli used for testing the system were created combining both global and local shifts. This is illustrated in figures 21-24, comparing the Praat (Boersma & Weenink, 2024) f_0 and intensity of the original recording to the VSG system output.

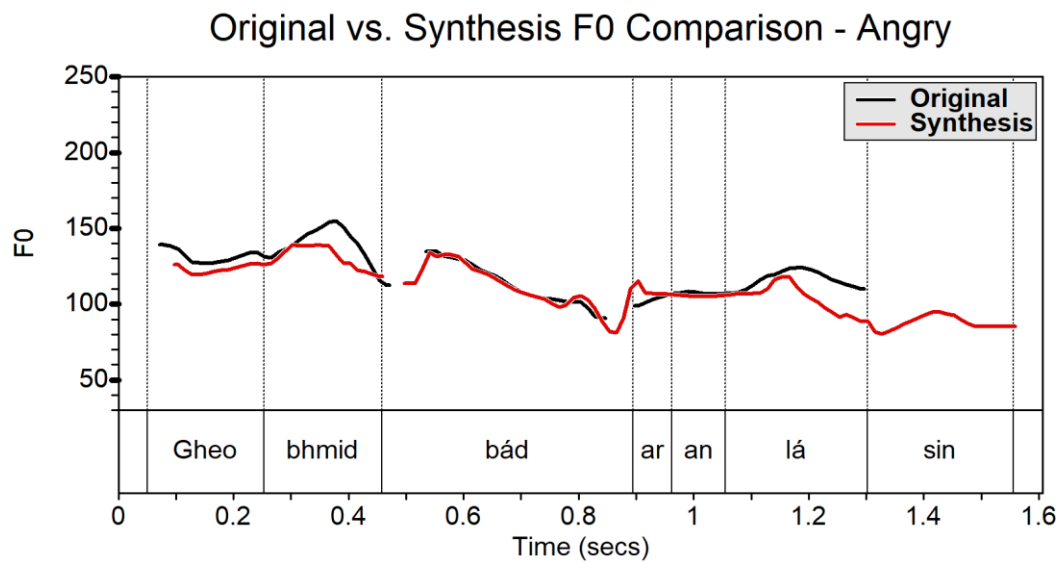


Figure 21 – Praat (Boersma & Weenink, 2024) f_0 values (in Hz) comparing the original recording and the VSG system output for the *angry* affect.

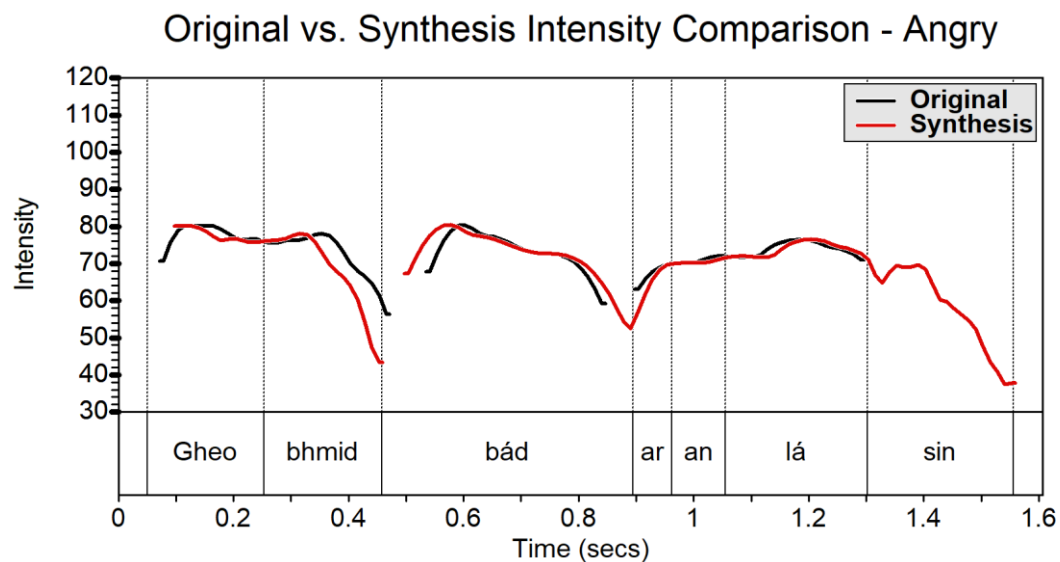


Figure 22 – Praat (Boersma & Weenink, 2024) intensity (in dB) values comparing the original recording and the VSG system output for the *angry* affect.

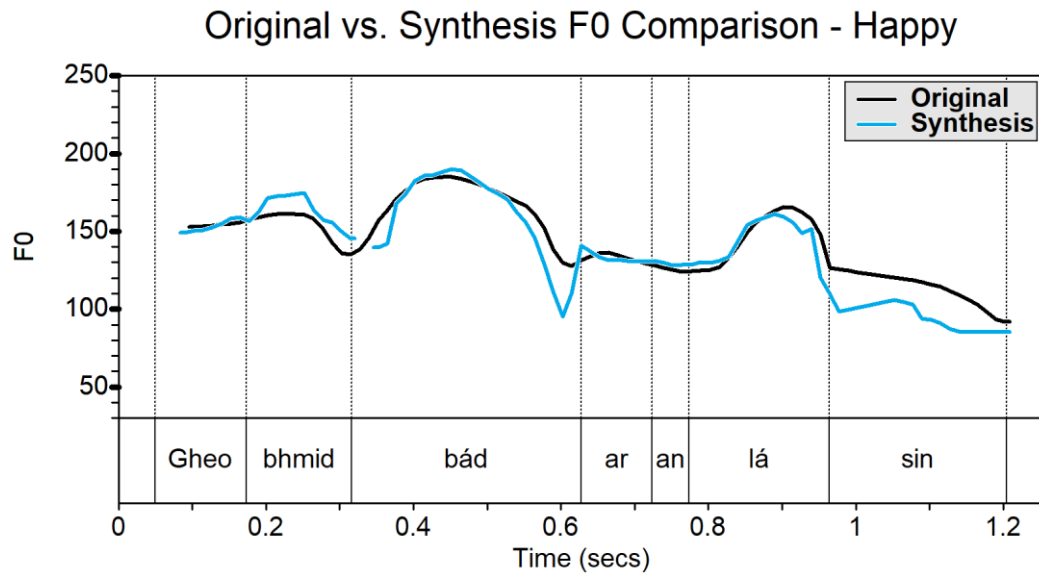


Figure 23 – Praat (Boersma & Weenink, 2024) f_0 values (in Hz) comparing the original recording (after the application of the Praat (Boersma & Weenink, 2024) f_0 adjustment to the second pitch accent) and the VSG system output for the *happy* affect.

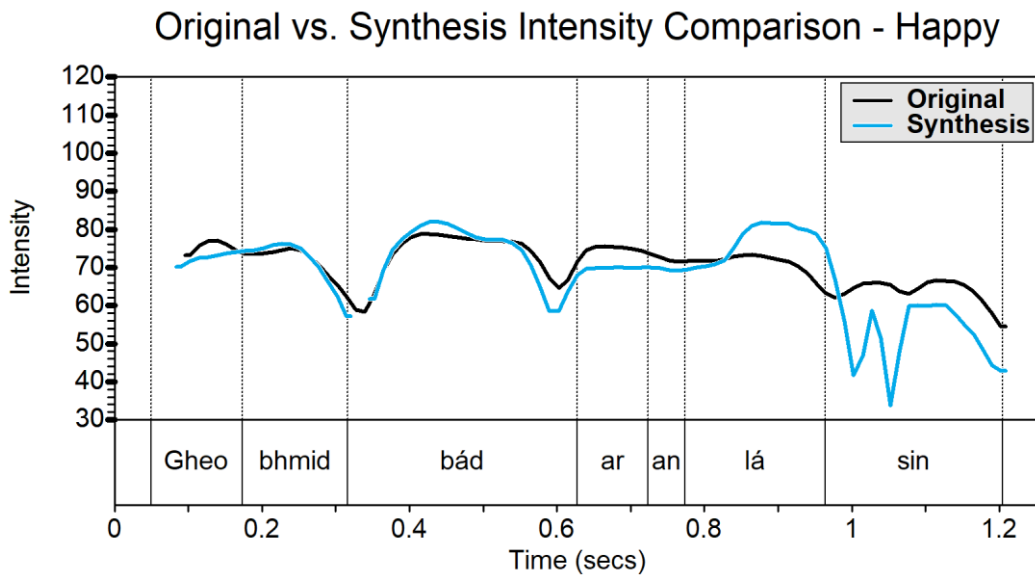


Figure 24 – Praat (Boersma & Weenink, 2024) intensity (in dB) values comparing the original recording and the VSG system output for the *happy* affect.

4.2.4. Global Manipulations

For the global transformations, the source parameters of the utterance are modified at the global level, i.e., a universal scaling formula is applied across the entire utterance. For this, a down-drift approach was used: a linear regression analysis is carried out on each modifiable voice source parameter on both the neutral and target (affective) utterances. Then, similar to the approach used in Haan et al. (1997), a secondary linear regression analysis (*Reg*) is conducted using only the data points either below (for f_0 and EE) or above (for RD) the overall regression line. This was done under the hypothesis that this analysis would be less affected by the local contours associated with the nuclear accents and therefore provide a better representation of the global voice source characteristics.

For f_0 and EE, the data points below the standard regression line were used, as for those parameter values are expected to increase during accented syllables even on the neutral baseline. It is the opposite for RD; lower RD values are expected on accented syllables as it inversely correlates to vocal fold tension. Thus, for RD, data points above the standard regression line were used to conduct the secondary regression analysis.

The resulting regression lines are used to calculate the global scaling factors. In addition, because the neutral baseline and the target utterance have different durations, timepoints at the segmental level are taken and inputted to the system to ensure the time-alignment of the regression lines. For each segmental time point of the neutral baseline, the scaling factor is calculated as follows:

$$S_{\text{global}}(t_{\text{neutral}}) = \frac{Reg_{\text{target}}(t_{\text{target}})}{Reg_{\text{neutral}}(t_{\text{neutral}})}$$

The scaling factors S at each time point are then linearly interpolated to compensate for the different number of time points between the neutral and target sentence, and to get the scaling factors of each pulse in the utterance. The baseline parameters are multiplied by the scaling factors, and new source parameters are obtained for resynthesis. Examples of the global manipulation output for the *angry* affect are presented in figures 25-

27 below; furthermore, all stylization images for all manipulation types can be found in Appendix B.

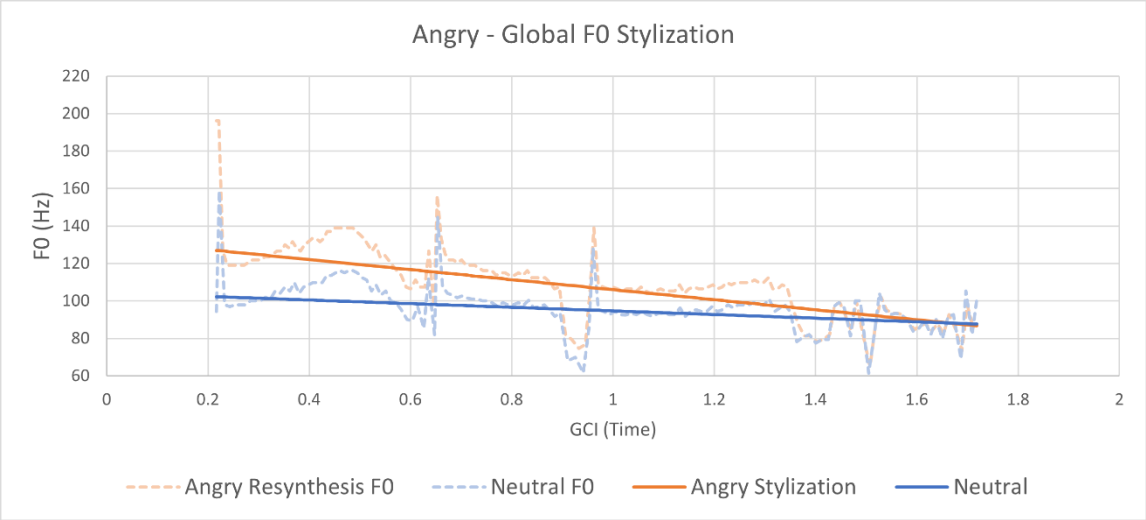


Figure 25 - Regression line f_0 stylization for the *angry* affect, overlaid onto the neutral analysis and angry global resynthesis contours.

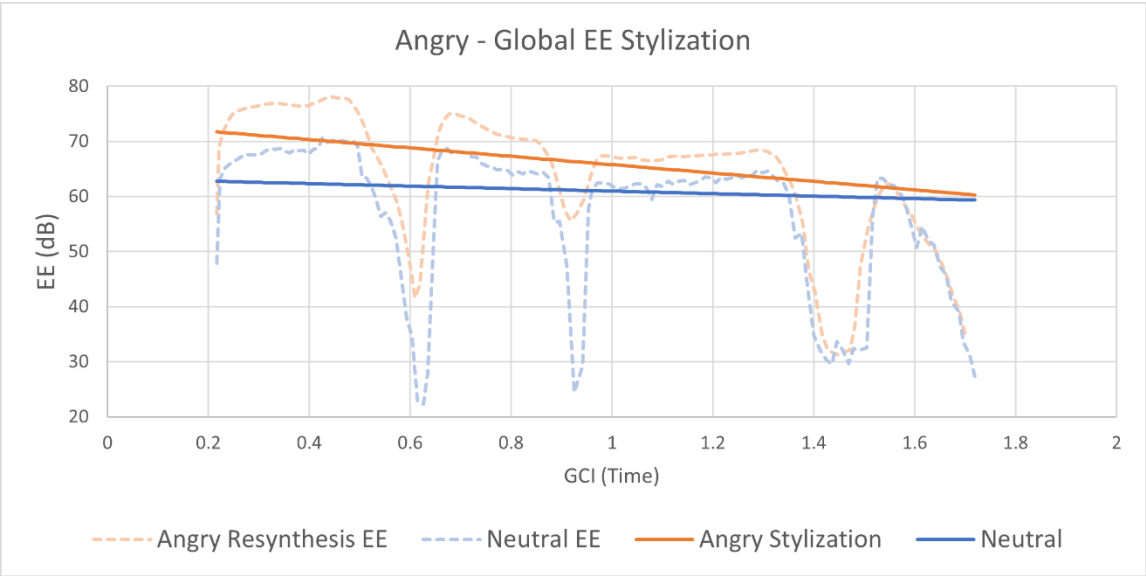


Figure 26 - Regression line EE stylization for the *angry* affect, overlaid onto the neutral analysis and angry global resynthesis contours.

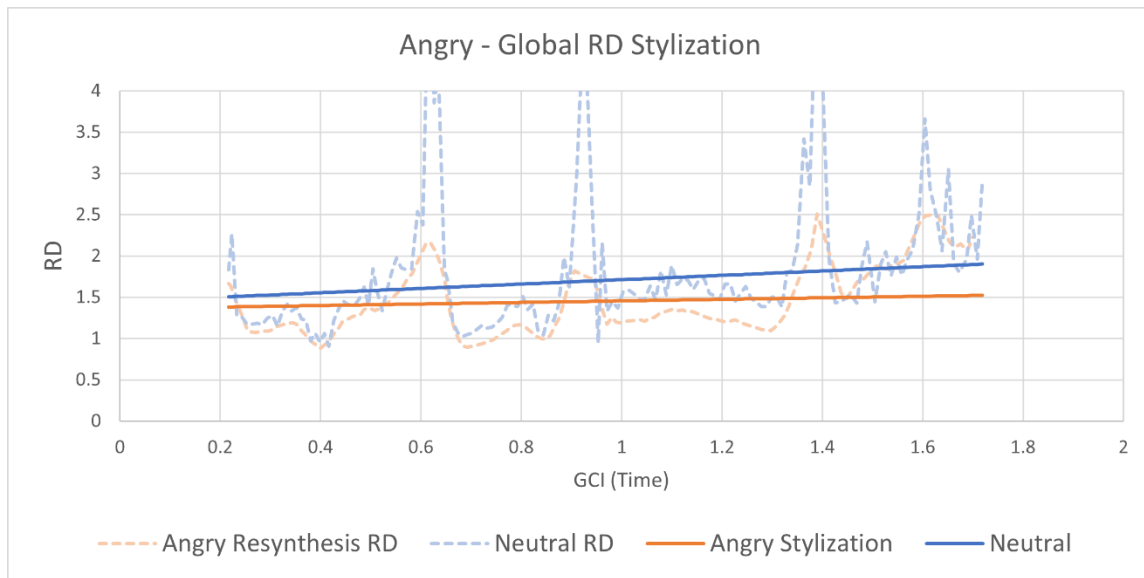


Figure 27 - Regression line RD stylization for the *angry* affect, overlaid onto the neutral analysis and angry global resynthesis contours.

4.2.5. Local Manipulations

The local transformations aim to impose stylized features of the target utterance's accented syllables onto the neutral utterance's accented syllables. For these, a simple anchor point system was used to obtain time points, extract measurements, and implement modifications. Those anchor points were based on phonetic level segmentations carried out in Praat (Boersma & Weenink, 2024). They are: (I) the first and last voiced pulses of the vowel in the accented syllables (*'bád'* and *'láj'*), unaltered in the local transforms and used only as parameter anchors; (II) pulses located 20% into the vowel from the anchor points in (I); and (III) the middle point of each accented vowel, i.e., the pulse at the midpoint between the anchor points in (I). Additionally, the local contours were stylized to create straight paths between anchor points. This was done by linear interpolation of the parameter values between these anchor points.

The local scaling factors were calculated for each parameter by dividing the interpolated parameter values by the corresponding values of the secondary regression line, i.e., the local shifts are the relative change of the target parameter values at the described anchor points from the target affect's regression line. These scaling factors are then transplanted onto the neutral secondary regression line for the local shifts – in other words,

local transforms are achieved by multiplying the scaling factors by the corresponding value from the secondary regression line of the neutral utterance, as follows:

$$S_{\text{local}} = \frac{P_{\text{local,target}}}{Reg_{\text{target}}}$$

Where P stands for the parameter value at the anchor points to be changed. Scaling factors at anchor points (I) described above are set to 1 to bring the contour from and back to the original neutral utterance. Only anchor points (II) and (III) in the neutral utterance's regression line are multiplied by the calculated scaling factors. With this, new source parameters are obtained, which are then used in the resynthesis:

$$P_{\text{local,transformed}} = S_{\text{local}} \cdot Reg_{\text{neutral}}$$

Examples of the local manipulation output for the *angry* affect are presented in figures 28-30 below.

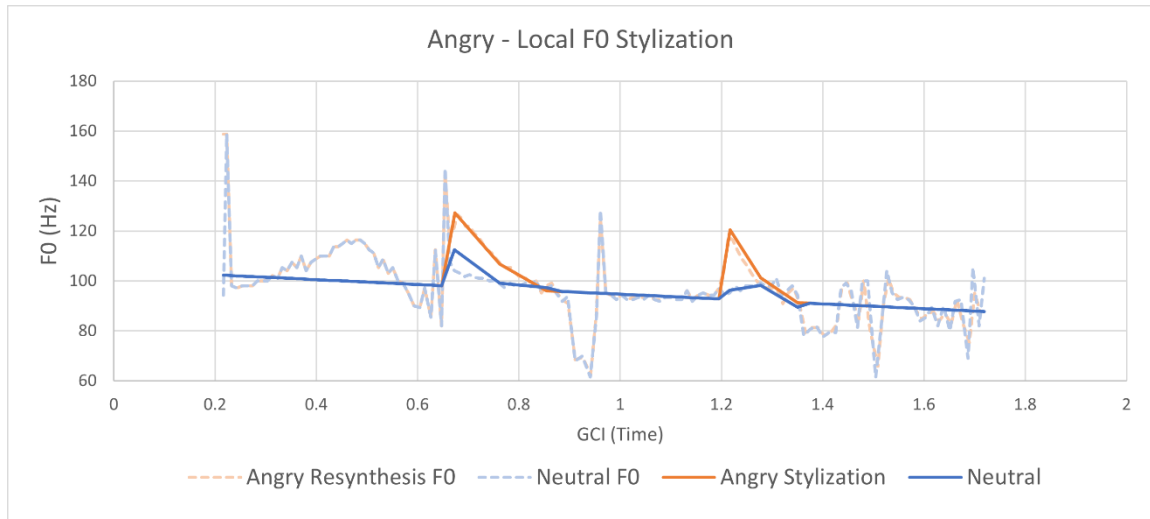


Figure 28 – Anchor point-based local f_0 stylization for the *angry* affect, overlaid onto the neutral analysis and angry global resynthesis contours.

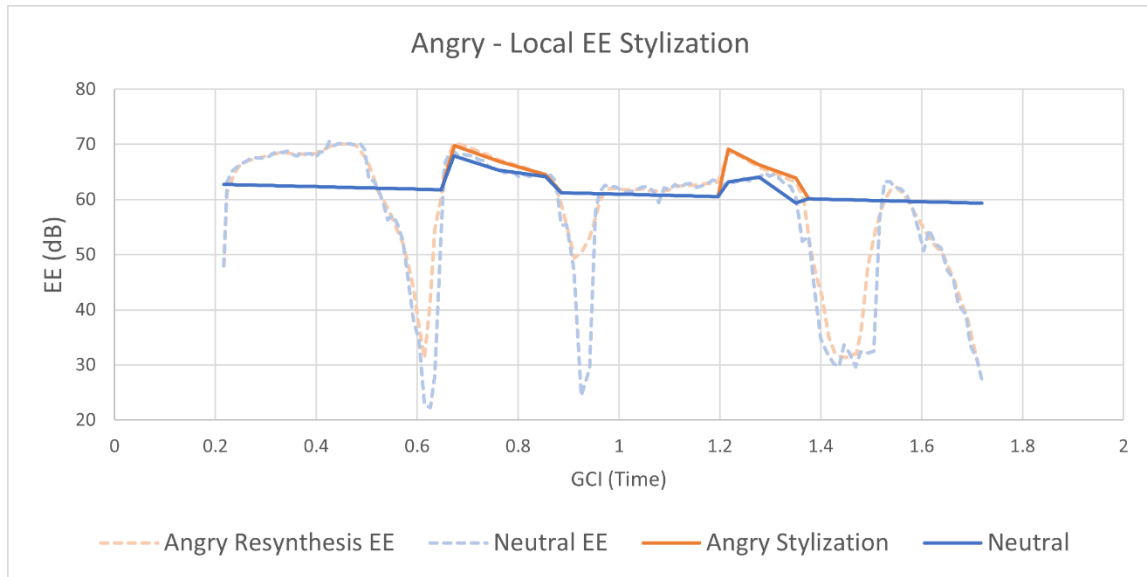


Figure 29 - Anchor point-based local EE stylization for the *angry* affect, overlaid onto the neutral analysis and angry global resynthesis contours.

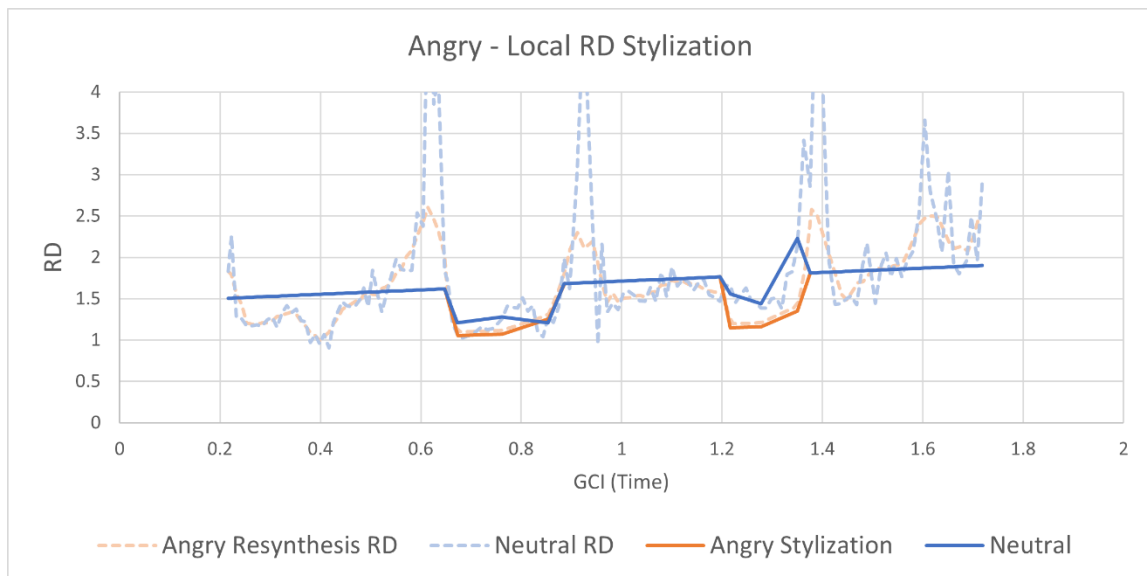


Figure 30 - Anchor point-based local RD stylization for the *angry* affect, overlaid onto the neutral analysis and angry global resynthesis contours.

4.2.6. Combined Manipulations

The combined stimuli bring together the global and local shifts with the scaling factors of the local changes applied directly to the secondary regression line of the target affect after the global shift is applied. In other words, the transformation of non-accented syllables is conducted in the same way as the global transformations, whereas the local transformation for accented syllables differs where the local scaling factors are multiplied onto the regression line of the target utterance after the global shift is applied, rather than on the original neutral:

$$P_{\text{combined,transformed}}(t) = \begin{cases} S_{\text{global}}(t) \cdot P_{\text{global,neutral}}(t) & t \text{ in non-accented region} \\ S_{\text{local}}(t) \cdot \text{Reg}_{\text{original}}(t) & t \text{ in accented region} \end{cases}$$

The new source signal is then generated using the modified source parameters. Examples of the combined manipulation output showing the combination of the regression line and anchor point stylizations for the *angry* affect are presented in figures 31-33.

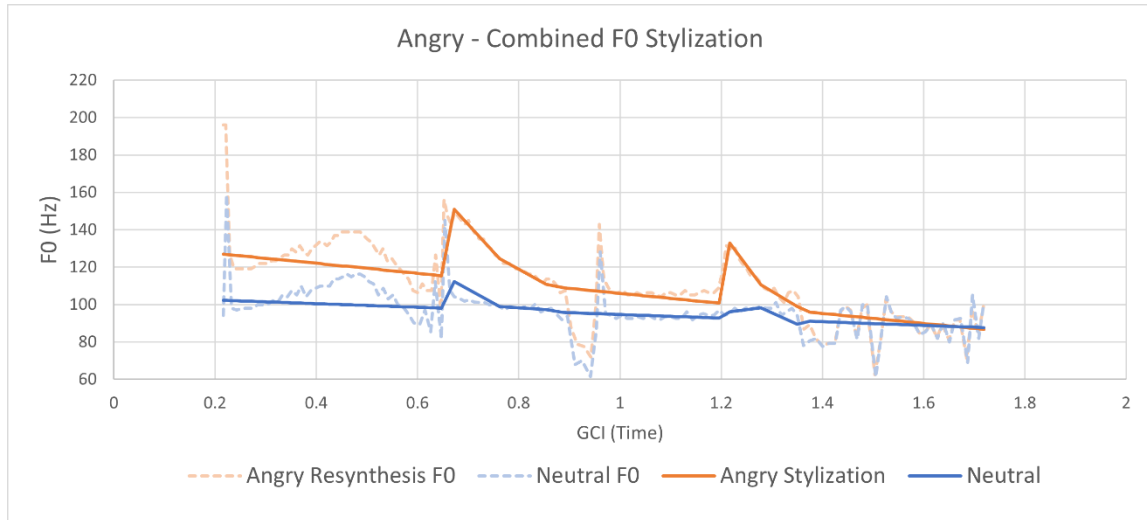


Figure 31 – Combination of the regression line and anchor point-based f_0 combined stylization for the *angry* affect, overlaid onto the neutral analysis and *angry* global resynthesis contours.

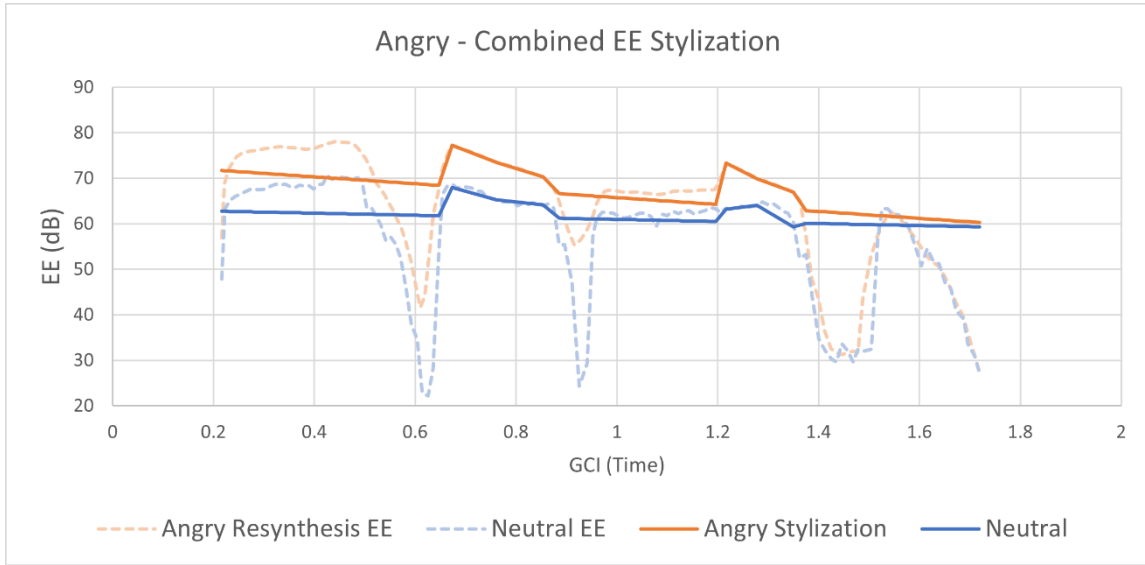


Figure 32 - Combination of the regression line and anchor point-based EE combined stylization for the *angry* affect, overlaid onto the neutral analysis and *angry* global resynthesis contours.

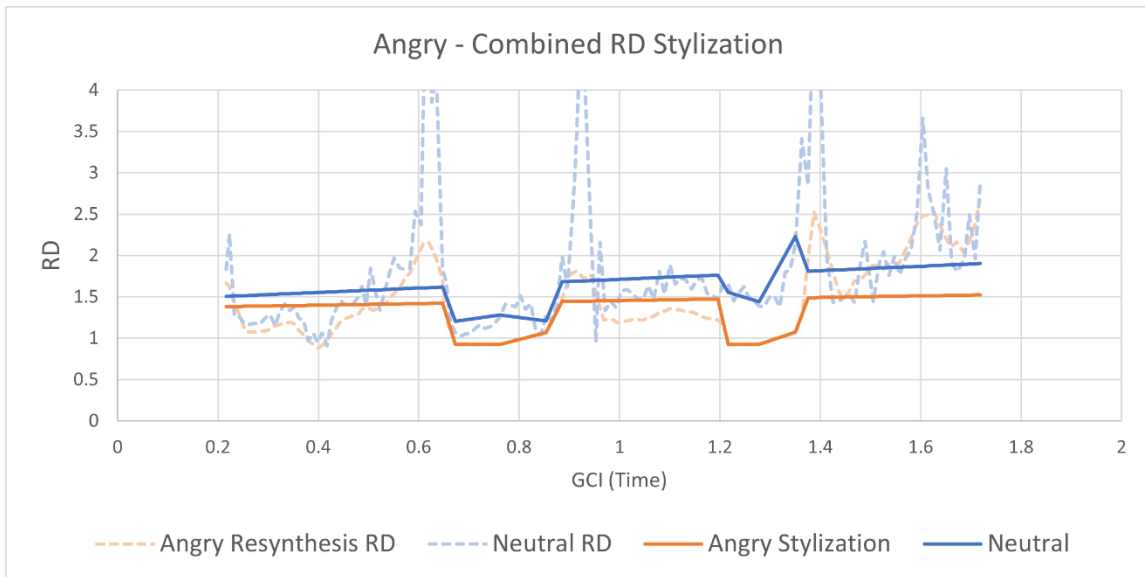


Figure 33 - Combination of the regression line and anchor point-based RD combined stylization for the *angry* affect, overlaid onto the neutral analysis and *angry* global resynthesis contours.

Because f_0 is changed, the new GCIs (glottal closure instants, or the time instants when the vocal folds snap shut) differ from those of the original neutral baseline utterance,

and therefore need to be recalculated. While there is an alternative approach that modifies f_0 in the resynthesized speech using the PSOLA algorithm (Moulines & Charpentier, 1990), in the VSG, PSOLA is only used to modify the f_0 of the source signal of the neutral input and to obtain the new GCIs to resynthesize the transformed speech, minimizing the distortion on the quality of the resynthesized speech.

With the new PSOLA-calculated GCIs on hand, the VSG system recalculates f_0 , and EE, RD and formant data are interpolated and resampled. Using the new f_0 , EE and RD values, the system then calculates all the other LF model parameters through the algorithm described in Z. Wang and Gobl (2023b). Finally, the new source signal is synthesized using the aliasing-free LF model described in Gobl (2021), and the transformed speech is generated putting this signal through a vocal tract model using formant resonators connected in cascade.

4.2.7. Other transformation details

In addition to the systematic changes described above, a few modifications were also made based on specific characteristics of the sentences used, issues with the recorded material, and current limitations of the system that required minor manual compensations. These changes were carried out mostly for the sake of the proposed experiments – any artefacts or oddities in the resynthesized material could potentially become a distractor for the listener judges. As the aim was for the focus of the listener to be on the effectiveness of voice source changes at conveying affect, stimuli therefore needed to be of the best audio quality achievable.

The final syllable ‘*sin*’ in sentence 2 caused issues in the regression analysis and parameter resynthesis due to the noisy nature of the fricative, the very short vowel and the following nasal at that point in the utterance. These caused extreme parameter value fluctuations, as well as artefacts whenever any kind of scaling was applied to that area. Because it was a sentence-final word mostly composed of noise with only a couple of very weak voiced pulses, this final syllable was excluded from the regression analysis and parameter values during the voiced portions of the syllable were always the same as for the neutral utterance.

As mentioned, the voice source is the main sound source for voiced sounds, i.e., vowels and voiced consonants. Therefore, standard inverse filtering of the data would not have provided measurements that allowed the resynthesis system to output unvoiced consonants. For that reason, dummy pulses were added to the neutral analysis for input into the VSG system so that stops and fricatives would be present in the output.

For these utterances to be more natural sounding in the final resynthesis, the glottal flow residual, defined as the difference between the output of the inverse filtering (i.e., the glottal flow derivative) and the modelled glottal flow derivative of the signal (Z. Wang & Gobl, 2022), was re-added to these areas to restore some of its original information. For the system to do this automatically, the unvoiced regions were here defined as the areas in the utterance that presented EE values lower than 1000. This also meant that the voiced regions would not have any residual added to them.

It bears mentioning that these dummy pulses were excluded when conducting the regression analysis for the global shifts, as their low EE values would skew the regression analysis. A MATLAB script⁵ was used to remove any data points with EE values below 500 before calculating the regression values.

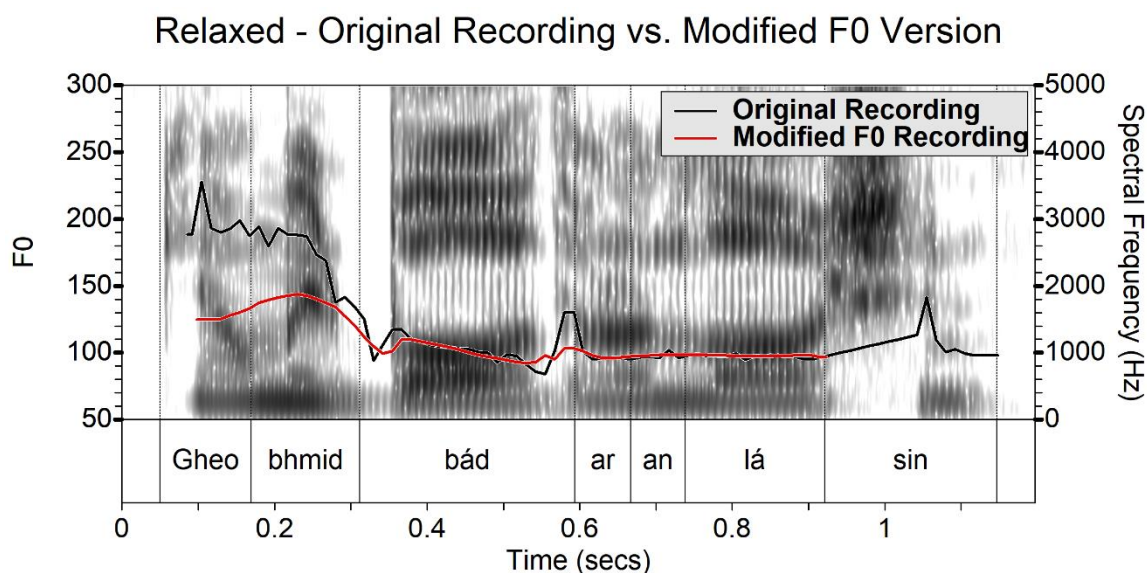


Figure 34 - F_0 contours between the original recording of the *relaxed* affect, and a version with modified f_0 .

⁵ Script written by Zihan Wang, author of the VSG system (Z. Wang & Gobl, 2023b).

For the *relaxed* stimulus, an additional manipulation was done to f_0 in Praat (Boersma & Weenink, 2024) (illustrated in figure 34). This was because the high, utterance-initial f_0 level on the pre-focal material was causing the regression analysis/declination slope to be extremely steep, which resulted in very low f_0 at the end of the sentence (i.e., much lower than the original production data level at that part of the sentence) giving the voice a creaky quality. Therefore, f_0 was lowered at the start of the utterance for *relaxed* and a new regression analysis was carried out for the application of global and combined shifts.

5. PERCEPTION EXPERIMENT 1 – RECORDED NEUTRAL BASELINE

Chapter 5 presents the first of the main perception experiments. The aim was to test simple voice transforms for cueing different affects. The transforms tested explore the relative effectiveness of three possible implementations, involving global, local and combined (incorporating both local and global) transformations of a neutral utterance. The parameter manipulations were guided by the detailed analysis of sentences portraying these emotions (presented in chapter 4) and past studies on the topic.

A set of stimuli was created according to the detailed methodology using the VSG system (Z. Wang & Gobl, 2023b). The neutral baseline used as input to the VSG system was the manual voice source analysis of the recorded neutral utterance '*Geobhaimid bá d ar an lá sin*', resynthesized by the VSG system through the inputted manual analysis parameters. This recorded neutral baseline was also used in the first set of visualizations that helped guide the stylization of parameter contours, also presented in chapter 4.

5.1. Modification Details

A total of 15 affect-targeting stimuli were created (global, local, and combined versions of the affects *happy*, *angry*, *sad*, *bored* and *relaxed*), for a total of 16 stimuli including the neutral baseline. Tables 3 and 4 present the global and local shifts, with parameter slopes and local contours illustrated in figures 35-40 and 41-46 respectively.

Table 3 - Global shift values (coefficient and intercept) for each affect and parameters (low regression line for f_0 and EE; high regression line for RD).

Affect	Parameters					
	F_0		EE		RD	
	Coefficient	Intercept	Coefficient	Intercept	Coefficient	Intercept
Neutral	-13.9	108.2	-2.7	63.2	0.3	1.5
Angry	-24.4	132.1	-6.9	72.8	0.0	1.4
Happy	-21.9	137.4	0.8	64.6	-0.2	1.8
Bored	-28.6	100.3	-4.1	59.5	0.1	1.7
Sad	-10.1	97.6	-7.3	64.7	0.3	1.7
Relaxed	-32.5	121.8	-4.5	63.9	-0.6	2.4

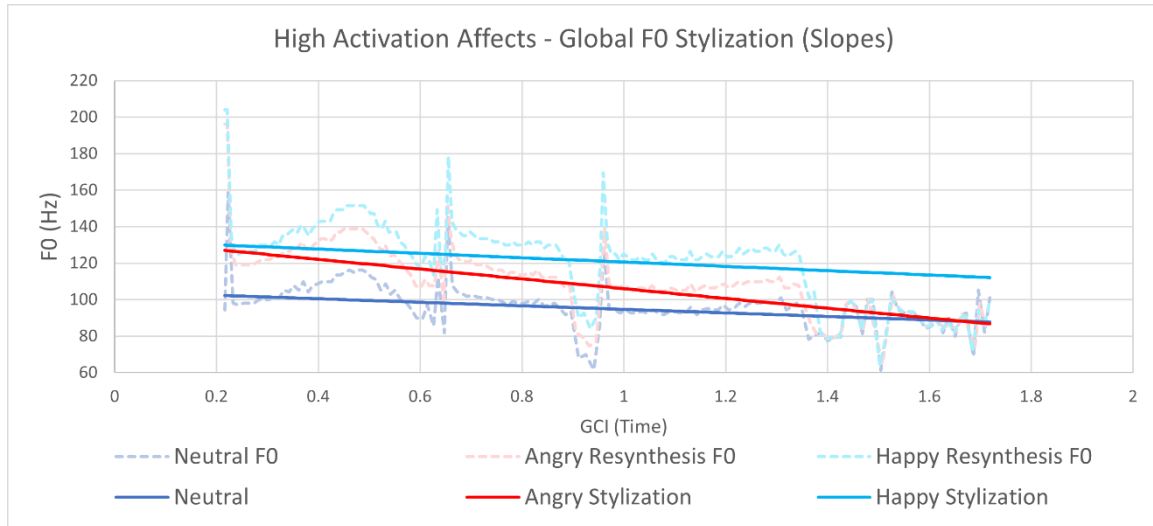


Figure 35 – Global slopes of high activation affects for f_0 , from values in table 3.

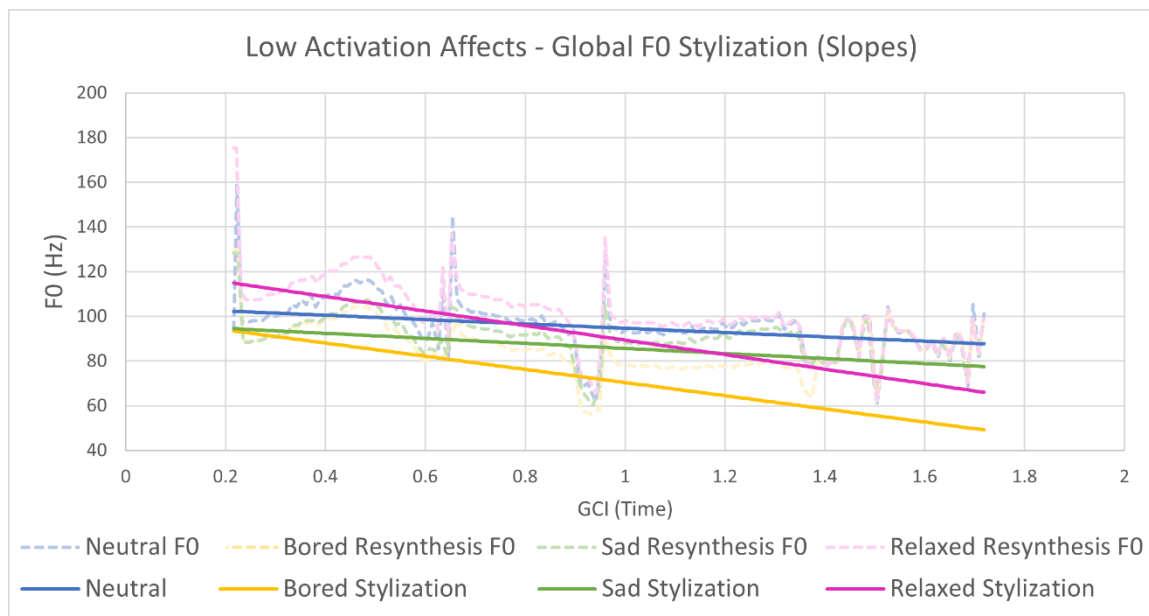


Figure 36 – Global slopes for the low activation affects for f_0 , from values in table 3.

Regression coefficients and intercepts for the global shifts are presented in table 3, including neutral for comparison. Figures 35 and 36 illustrate the regression slopes for f_0 , split between activation categories for readability. As shown, *angry* features a steeper slope than neutral while *happy* has a comparatively softer slope with a higher starting point, being the highest global rise in f_0 across the presented affects. *Bored* and *relaxed* also show steeper slopes than neutral, but with lower and higher intercepts for *bored* and *relaxed*, respectively. *Sad*, as expected, has a flatter contour and overall lower f_0 .

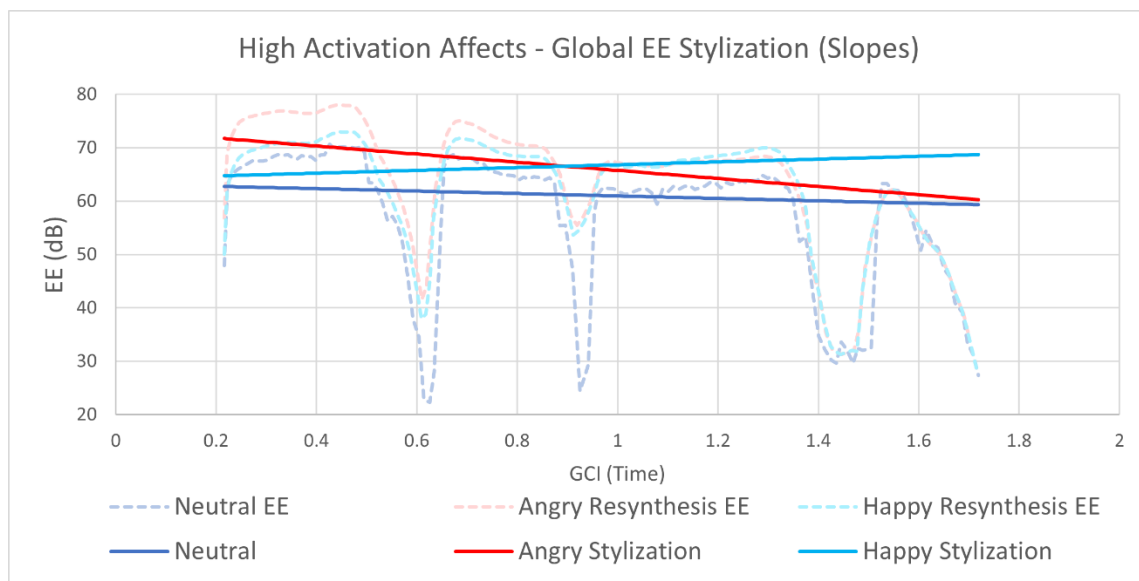


Figure 37 - Global slopes of high activation affects for EE, from values in table 3.

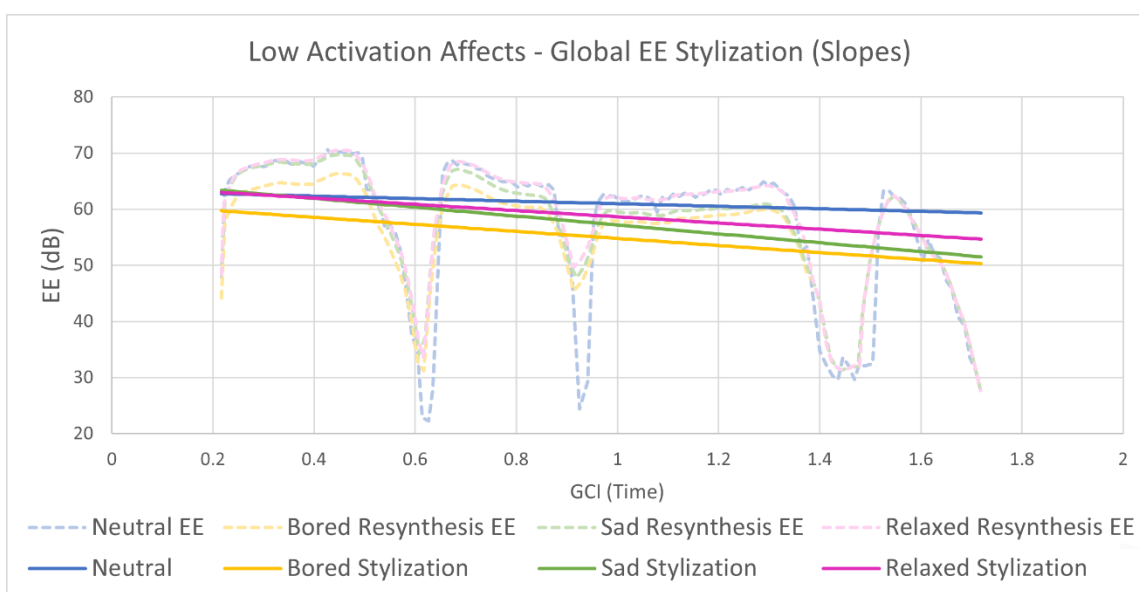


Figure 38 - Global slopes of low activation affects for EE, from values in table 3.

EE regression slopes are shown in figures 37-38. EE shows a similar trend to f_0 for *angry*, but *happy* has a rising slope due to the accented syllables. *Bored* and *relaxed* again show a steeper slope than neutral, but with lower and higher intercepts, respectively. *Sad*, in contrast, has a much more dramatic drop than neutral, reflected in the intercept of the

regression in contrast to f_0 and likely due to the voice becoming quieter through the utterance, which aligns with descriptions of this affect.

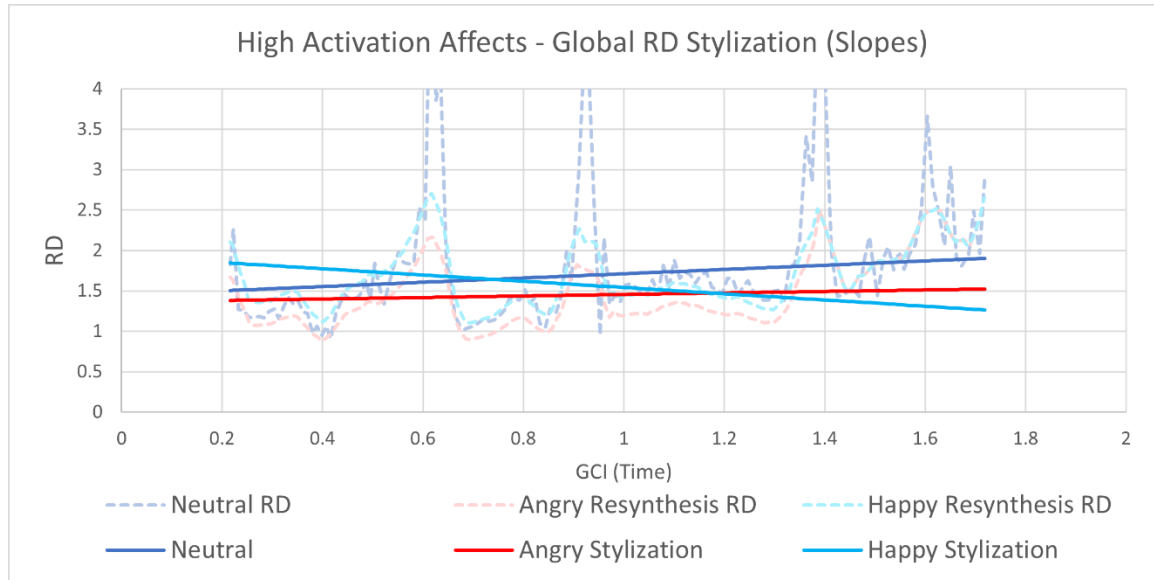


Figure 39 - Global slopes of high activation affects for RD, from values in table 3.

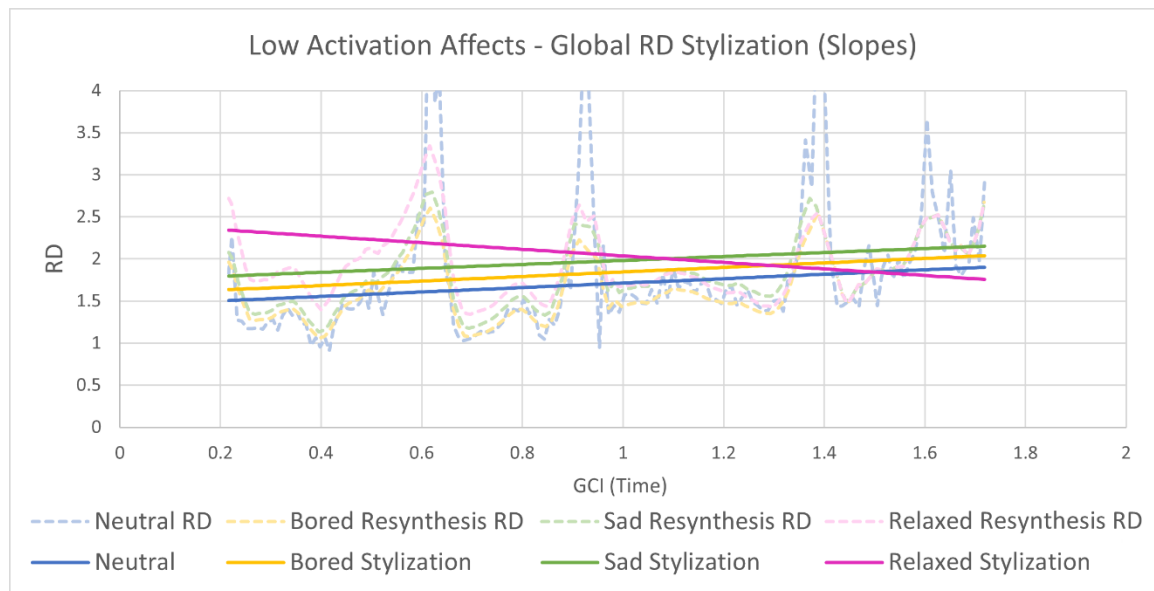


Figure 40 - Global slopes of low activation affects for RD, from values in table 3.

As for RD, shown in figures 39-40, neutral progressively becomes laxer through the utterance, while *angry* shows a flatter slope and a lower intercept indicating a tenser nature. *Happy* shows this even more evidently, with a higher intercept. *Bored* and *sad* both show a

higher, laxer start, with *bored* having a steeper slope than neutral indicating higher tension, and *sad* having the same coefficient as neutral, being slightly globally laxer. *Relaxed*, by definition, starts at a much higher intercept than neutral, and shows a much more dramatic, contrasting rising slope.

Table 4 lists the local scaling factors (i.e., multipliers applied to the neutral) for each affect, applied according to the anchor points detailed in chapter 4. These represent the deviation of the anchor points from the overall lower regression line, i.e., the VSG (Z. Wang & Gobl, 2023b) takes the target values of each anchor point for each recorded affect sample, and calculates the scaling factors between these values and the regression lines they are applied to. Namely, they consist of a pair of anchor points 20% into the accented vowel from its starting boundary (II₁) and 20% into the vowel from its ending boundary (II₂), as well as the anchor point in the middle of the vowel (III).

These local contours are also illustrated in figures 41-46 including both the resynthesis contour and its stylized form based on the anchor points below, with the neutral regression line serving as reference for the non-focal material in the latter case.

Table 4 - Local shift values for each affect, applied as scaling factors over the neutral or target regression line through the anchor point system described in chapter 4. II₁ and II₂ refer to the anchor points 20% into the accented vowels (*bád* and *lá*), from the starting boundary of the vowel and to the ending boundary of the vowel, respectively. III refers to the anchor point in the middle of the vowel, as detailed in section 4.2.

Affect	Parameters & Anchor Points								
	F0			EE			RD		
	II ₁	III	II ₂	II ₁	III	II ₂	II ₁	III	II ₂
Angry	1.3	1.1	1.0	2.8	2.0	1.5	0.7	0.7	0.8
Happy	1.4	1.5	1.3	2.4	2.4	2.2	0.7	0.8	0.7
Bored	1.1	1.1	1.0	1.5	1.2	0.9	0.8	0.9	1.0
Sad	1.1	1.1	1.0	1.3	1.5	1.5	1.0	0.8	0.8
Relaxed	1.0	1.0	1.0	2.0	1.7	1.3	0.9	0.8	1.0

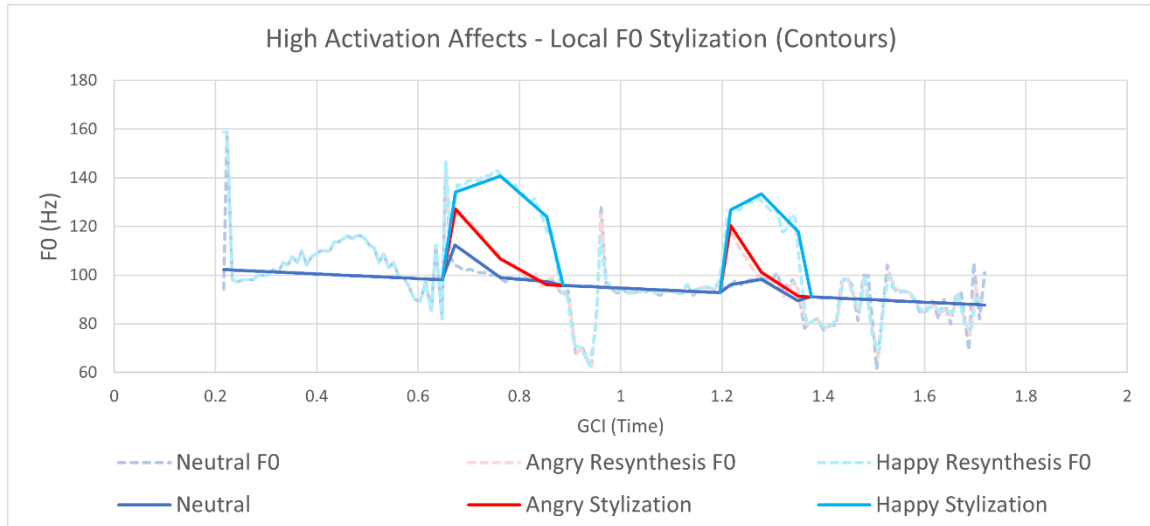


Figure 41 – Local contours of high activation affects for f_0 , from values in table 4.

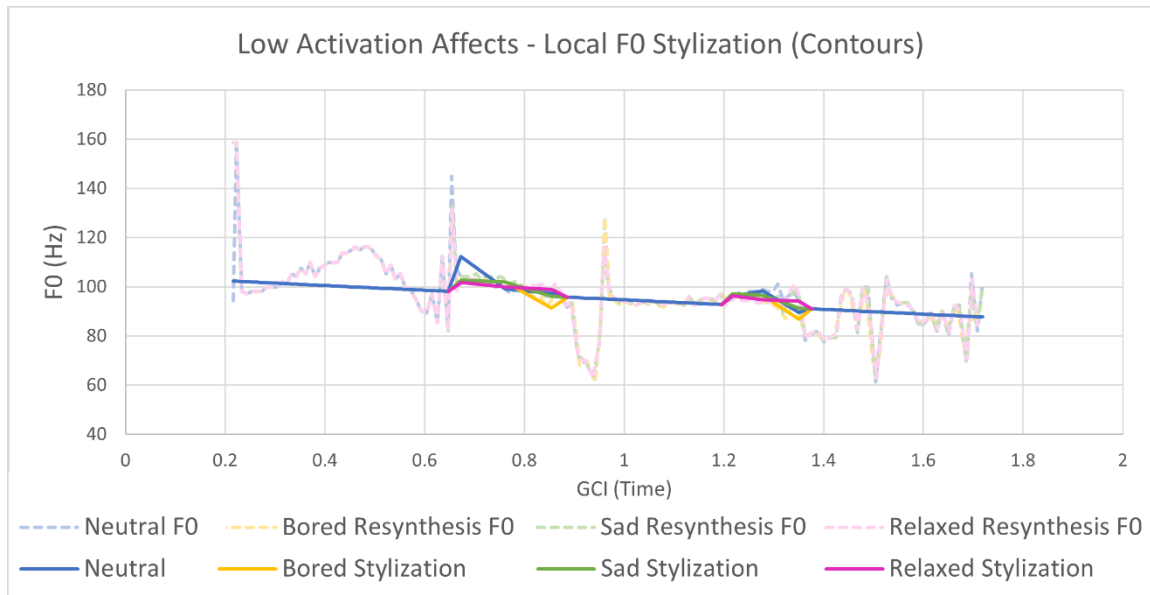


Figure 42 – Local contours of low activation affects for f_0 , from values in table 4.

For f_0 , shown in figures 41-42, *angry* raises the start of the vowel, producing a subsequent triangular drop. *Happy* has the high, slightly rounded contour seen previously. *Relaxed* stays close to neutral in terms of f_0 contour, with slightly higher anchor point values. *Bored* and *sad* show little difference from the neutral, with minimal rises at the start of the vowel and, for *bored*, a drop below neutral. While this goes against the evidence in the literature for the *sad* affect being flatter in contour, one must consider that this scaling is

affected by the simplification of the neutral and the *sad* contours as their lower regression analyses, from which these factors are calculated and applied to.

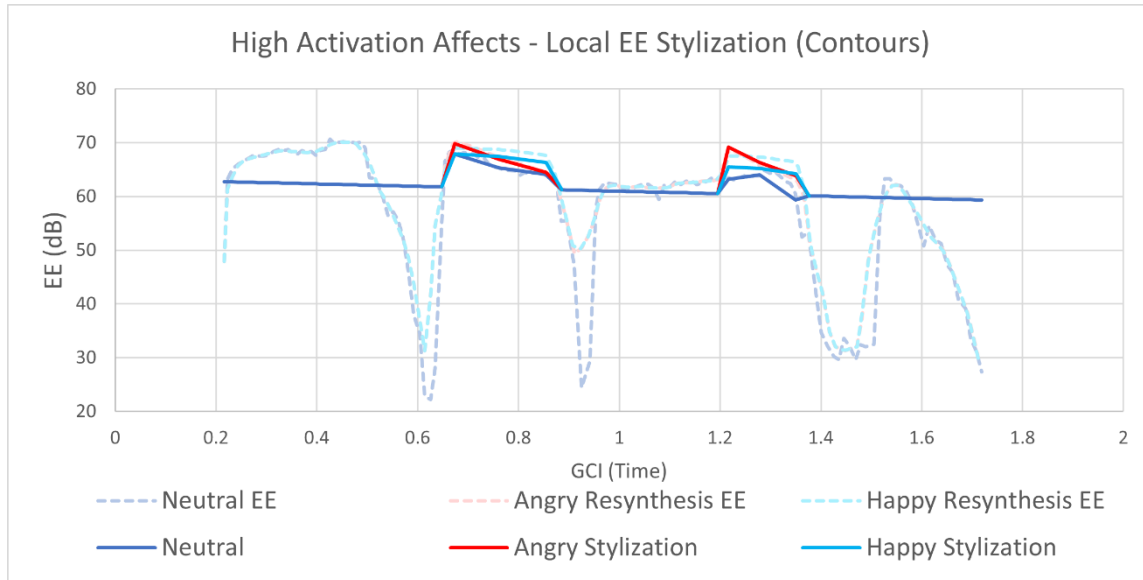


Figure 43 - Local contours of high activation affects for EE, from values in table 4.

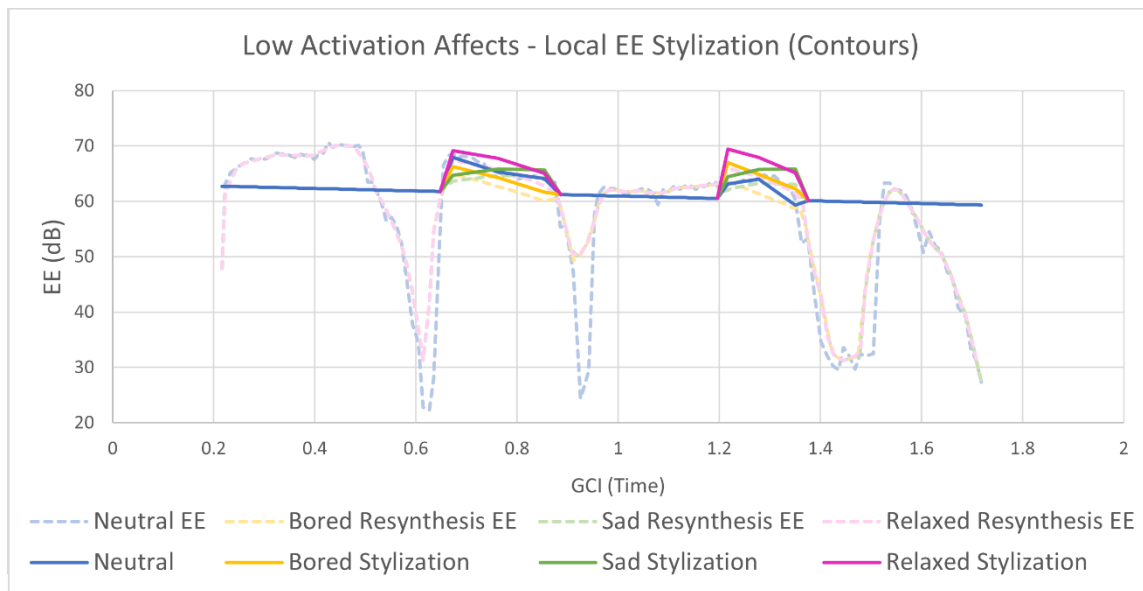


Figure 44 - Local contours of low activation affects for EE, from values in table 4.

EE (figures 43-44) for *angry* reflects f_0 , showing a steep rise and drop in the parameter levels through the accented syllable. *Happy* is also similar to f_0 with a stable, high contour. The same can be said for *bored* where it starts higher and then drops slightly below

the neutral. Here, *sad* shows the expected contour – while it is higher than neutral based on the relative shift of the pitch accents from the baseline, the contour is flattened – while neutral shows a higher start and a gentler fall, the first anchor point of *sad* is lower than its following ones, creating a rising contour. Finally, *relaxed* reflects the f_0 contour with a higher start and a noticeable, progressive drop back to neutral.

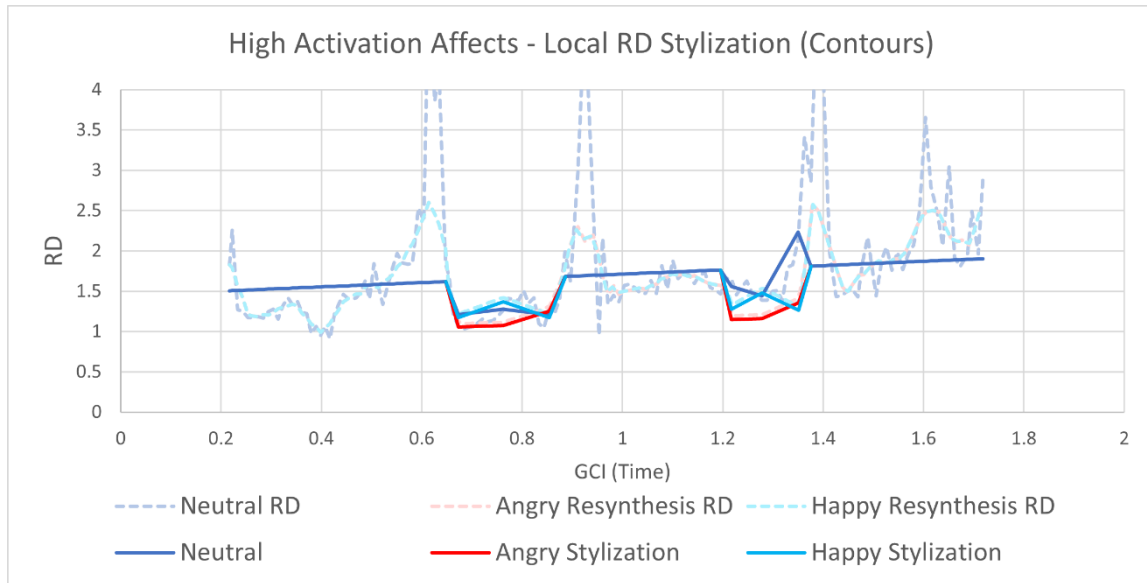


Figure 45 - Local contours of high activation affects for RD, from values in table 4.

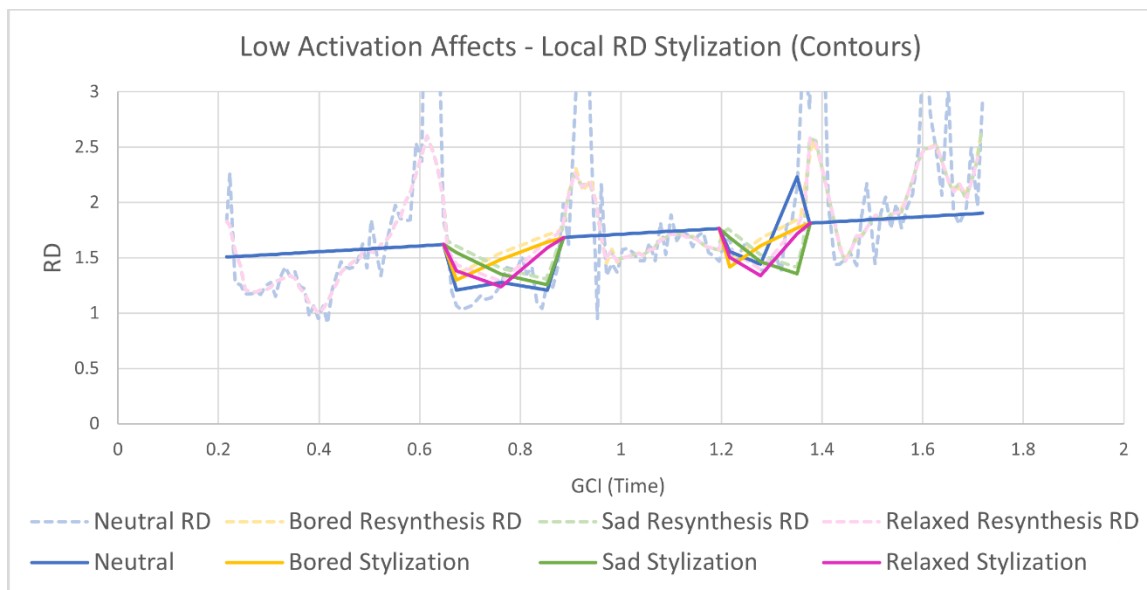


Figure 46 - Local contours of low activation affects for RD, from values in table 4.

As for RD (figures 45-46), *angry* starts considerably tenser and features a slight drop, still maintaining its expected tense voice relative to neutral. *Happy* also shows increased tension, with a contour similar to f_0 . *Bored*, *sad* and *relaxed* also present overall slightly tenser levels than neutral, but the same argument for f_0 in *sad* applies – these factors are calculated and applied relative to the regression analysis, and don't necessarily reflect the detailed contours shown in the figures presented in chapter 4.

It bears mentioning once more that these scaling factors are approximations of the manual analysis of data. They are a compromise between the local measurements of the two accents for the sake of simplification rather than a direct reflection of the analysis values, applied equally to both pitch accents in the neutral baseline. For reference, figures 47-49 below⁶ show these shifts applied by the VSG system (Z. Wang & Gobl, 2023b) to create the *angry combined* stimulus, with the global shifts, defined by the lower regression line, applied and followed by the relative local shifts described by the anchor point system onto this new regression line. Similar illustrations for all the other manipulations and affects can be found in Appendix C.

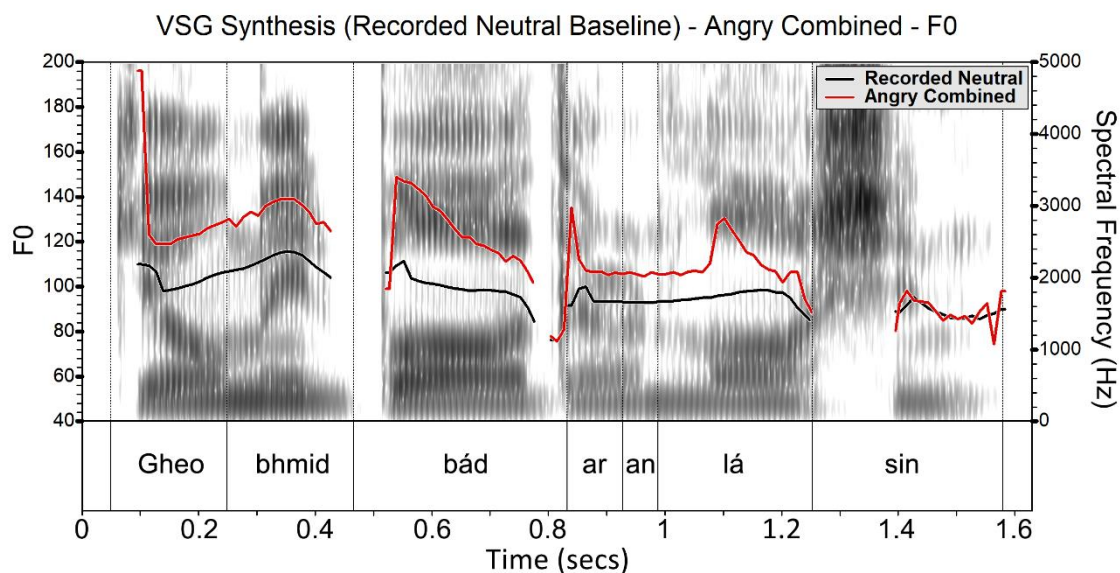


Figure 47 - F_0 contour of the combined modification of the recorded *neutral* stimulus into the *angry* target affect.

⁶ Parameter illustrations in figures 47-49 and in Appendix C were created through the custom Praat script by Rodgers (2022) mentioned previously and edited through Adobe Photoshop (Adobe Inc., 2023) for additional formatting adjustments.

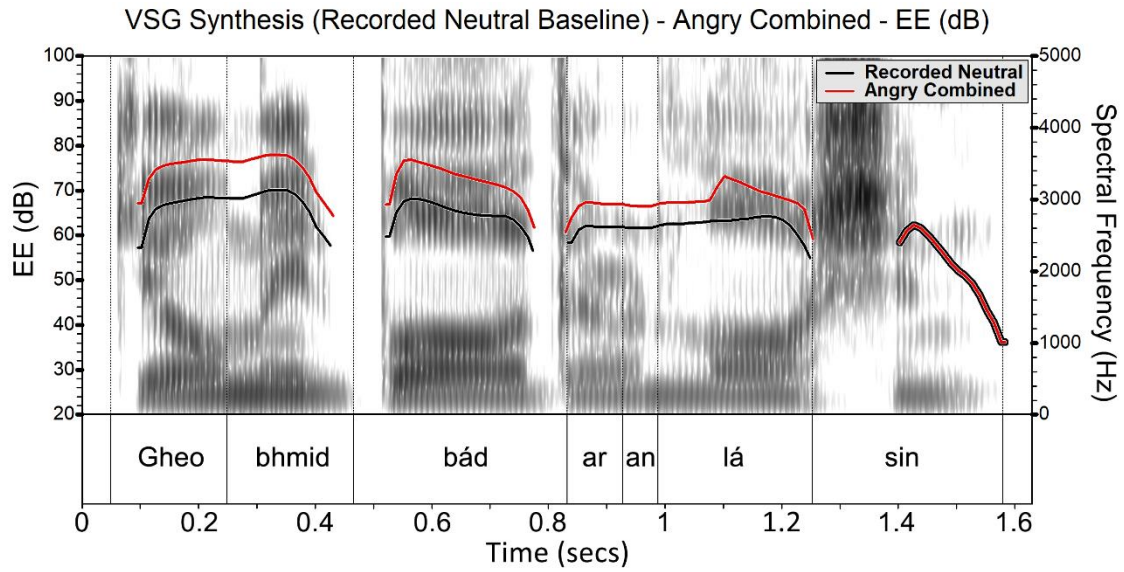


Figure 48 - EE contour of the combined modification of the recorded neutral stimulus into the *angry* target affect.

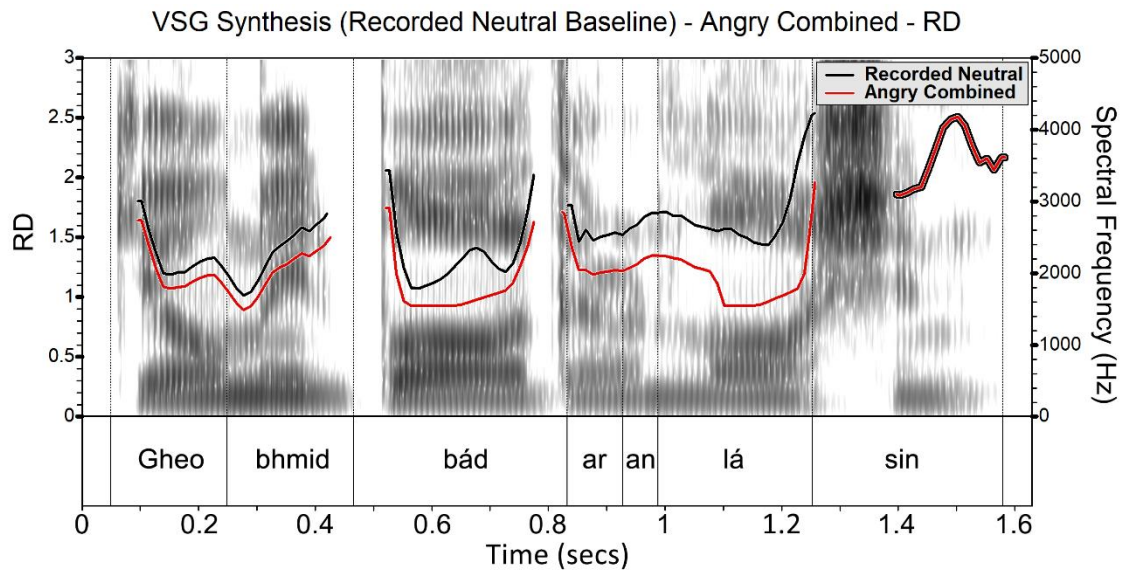


Figure 49 - RD contour of the combined modification of the recorded neutral stimulus into the *angry* target affect.

As we can see, the resulting combined stimulus follows expectations from the literature, and reflects the analysis illustrated in section 4.1 taking the proposed simplification of contours into consideration. Lastly, tables 5-7 show the modifications in terms of mean difference values between the neutral resynthesis of the VSG and the target affect output, for all modification types. These measurements were once again taken to

check the accuracy of the system's output according to the desired result. F_0 measures are shown in semi-tones, EE in dB, and RD on a linear scale.

Table 5 - Mean Difference values taken across the full utterance between the affective stimuli and the neutral baseline. Highlighted values are those most relevant to evaluate the performance of the VSG system in applying the global, utterance-wide shifts.

Affect	Mean Difference Values – Global vs. Local vs. Combined (Full Utterance)								
	F0 (semi-tones)			EE (dB)			RD		
	Global	Local	Combined	Global	Local	Combined	Global	Local	Combined
Angry	2.4	0.4	2.8	6.0	0.9	7.3	0.9	1.0	0.8
Happy	4.1	1.7	6.0	4.1	1.6	5.7	1.0	1.0	1.0
Bored	-2.0	0.0	-2.1	-3.4	-0.5	-3.9	1.0	1.0	1.0
Sad	-1.0	0.1	-1.0	-1.3	-0.2	-1.6	1.1	1.0	1.1
Relaxed	1.0	0.1	0.9	0.4	0.3	0.6	1.2	1.0	1.2

Table 6 - Mean Difference values taken across the first pitch accent of the utterance *bád* (between anchor points II1 and II2 defined in table 4), between the affective stimuli and the neutral baseline. Highlighted values are those most relevant to evaluate the performance of the VSG system in applying the desired shifts (i.e., the local shifts applied specifically to the pitch accent *bád* in the local column, and the combined shifts applied to the same pitch accent but with as the addition of the local and global shifts).

Affect	Mean Difference Values – Global vs. Local vs. Combined (First Accent - Bád)								
	F0 (semi-tones)			EE (dB)			RD		
	Global	Local	Combined	Global	Local	Combined	Global	Local	Combined
Angry	2.7	1.4	4.1	6.4	1.5	8.0	0.8	1.0	0.8
Happy	4.9	4.6	9.5	4.0	2.6	6.3	1.0	1.0	1.1
Bored	-2.5	0.0	-2.7	-4.1	-2.7	-6.8	1.0	1.2	1.2
Sad	-1.3	0.2	-1.3	-1.7	-1.3	-3.0	1.1	1.1	1.3
Relaxed	1.1	0.1	1.1	0.2	-0.3	-0.1	1.2	1.1	1.4

Table 7 - Mean Difference values taken across the first pitch accent of the utterance *lá* (between anchor points II1 and II2 defined in table 4), between the affective stimuli and the neutral baseline, for additional reference.

Affect	Mean Difference Values – Global vs. Local vs. Combined (Second Accent - Lá)								
	F0 (semi-tones)			EE (dB)			RD		
	Global	Local	Combined	Global	Local	Combined	Global	Local	Combined
Angry	1.9	0.8	2.8	4.1	2.4	6.9	0.8	0.8	0.6
Happy	4.5	3.9	8.8	5.8	3.5	9.9	0.9	0.9	0.8
Bored	-3.8	-0.5	-4.4	-4.3	-1.3	-5.6	0.9	1.0	0.9
Sad	-0.7	-0.3	-1.2	-3.1	0.0	-3.4	1.1	1.0	1.0
Relaxed	0.2	-0.3	-0.2	-0.1	0.7	1.1	1.0	1.0	0.9

The values most relevant for evaluating the effectiveness of the VSG at applying the proposed manipulations can be found in the *global* columns of the top chart (reflecting the global changes across the whole utterance), and the *local* columns of the middle chart (reflecting the local changes in the first pitch accent), highlighted in yellow. Additionally, the *combined* columns should ideally be the addition of global and local shifts (or rather, sum for f_0 and EE values and multiplication for RD values), bar a few fluctuations due to PSOLA, stylization and rounding of values.

Values shown are sensible – *angry* and *happy* have increased f_0 and EE values for both global and locals shifts, especially the latter, and the combined values reflects these shifts. For RD, differences were small, but still reflected as *angry* has lower RD values (tenser voice). Low activation affects show a decrease in global f_0 and EE levels, but not as dramatic a change in local measures since, as mentioned, these affects were close to the neutral from the start. *Relaxed* additionally shows an expected rise in RD.

5.2. The Happy Affect Pre-Test

In addition to the general shifts described above, further modifications to the *happy* and *sad* stimuli were explored. This was because the initial impressions of the *happy* stimulus with only voice source shifts did not seem very effective in portraying its target, and because happiness has been a persistently elusive emotion to capture in terms of vocal parameters. Previous evidence has, however, shown happiness to be associated with a faster speech rate, as well as an increase in formant frequencies due to the change in the proportions of the vocal tract caused by the smiling action (Torre, 2014) which indicate that, for the *happy* affect, the filter plays a significant role.

A small listening experiment to test the effectiveness of simple, utterance-wide tempo and formant shifts to the *happy* stimuli was conducted as follows. A set of *happy*-targeting stimuli was created through the VSG, only modifying the voice source parameters to create global, local and combined stimuli as previously described. This was called the *Source Only* set. This set was then copied and further modified into three new sets, each with different additional manipulations conducted in Praat (Boersma & Weenink, 2024).

One set, called the *Tempo Shift* set, had its tempo altered through the Duration Manipulation in Praat (Boersma & Weenink, 2024) though the PSOLA algorithm (Moulines & Laroche, 1995). The tempo shift was calculated by dividing the total duration of the chosen recorded *happy* stimulus of sentence 4 (to account for the duration of the proper second accented syllable without a focus shift) by the duration of the recorded *neutral* stimulus of the same sentence. The obtained value resulted in a 25% tempo reduction that was applied to the *happy* stimuli from the VSG output, uniformly across the entirety of the utterances.

The second set of stimuli, the *Formant Shift* Set, had its formant frequencies altered through the Change Gender Conversion algorithm in Praat (Boersma & Weenink, 2024). A value of 4% was chosen, which is large enough to be clearly noticeable and small enough not to affect the perceived speaker identity. Finally, the third set of stimuli, the *Formant + Tempo* Set, had both its tempo and formants shifted by the same values, totaling 12 stimuli for testing by including the *Source Only* stimuli.

Stimuli were presented to a group of 10 listeners, all Irish speakers, in the form of a perception test. Each set of 12 stimuli was randomized and presented 5 times to listeners as a series of videos via Google Forms, and responses were collected on paper. Videos consisted of the neutral stimulus followed by the *happy*-targeting stimulus; listeners were then asked to what degree they found the second stimulus to conjure happiness in a scale from 0 (Neutral/No emotion) to 5 (Very Happy).

Results were then pooled and analyzed by the mean values of the perceived strength from the scale assigned to each stimulus and can be seen in figure 50 below. While the number of listeners was small, results show a consistent tendency towards the combination of formant and tempo shifts added to voice quality as the most effective at conveying happiness for all modification types. These shifts also do not seem to undermine the voice source shifts. Therefore, it was decided that formant and tempo shifts would indeed be applied to the stimuli of the larger perception test.

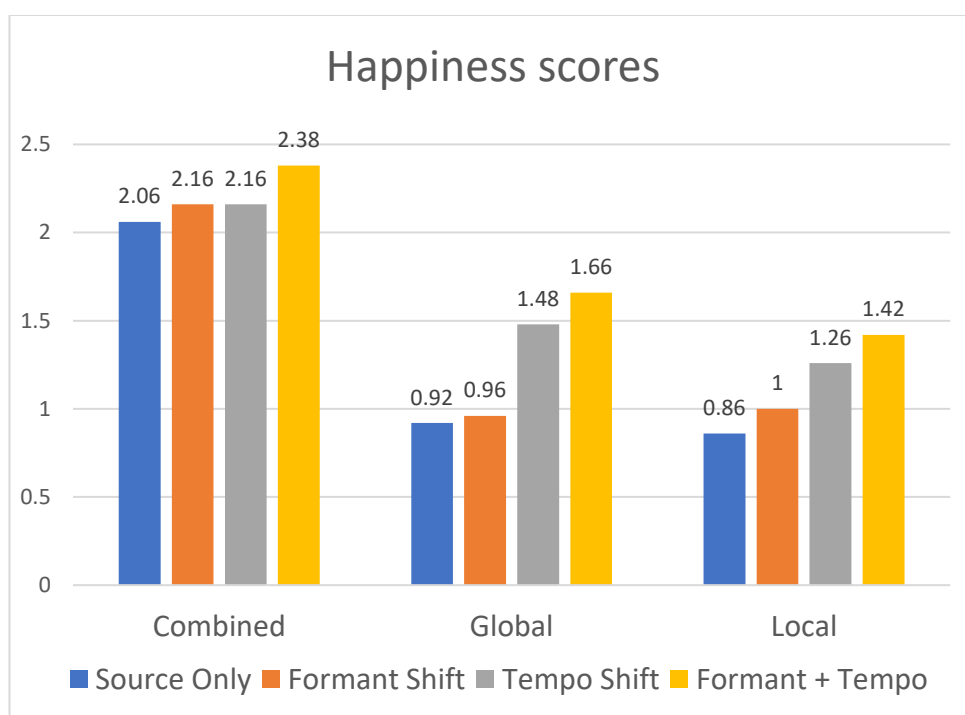


Figure 50 - Results from the Happiness test comparing stimuli without and with formant and/or tempo shifts, in terms of mean scores.

In addition, feedback was collected from colleagues that partook in the test saying that, while the voice did sound happier with the tempo shift, it was a little too extreme and made the utterance sound unnaturally sped up. Therefore, the tempo shift was reduced for the final set of stimuli.

5.2.1. Final Manipulation Details

Based on the pre-test detailed above, final *happy* stimuli used in this project's experiments had a 4% formant shift along with a 10% decrease in duration (i.e., utterances were made 10% faster). This 10% value was acquired through informal listening sessions and based on listener's feedback, and being considered more natural sounding than the original tempo shift.

In the case of the *sad* stimuli, modifiable voice parameter levels were too close to the *bored* stimuli, which caused them to only be differentiated in terms of very subtle contour differences in the accented syllables. This was also a result of there being no further voice

source modifications to the parameters other than the simplified system outlined above – creaky voice, for example, which is very often associated with boredom, was not included. Therefore, *sad* and *bored* were further differentiated in terms of global tempo, as, while a bored voice is often associated with both faster and slower speech rate, a sad voice is consistently associated only with slower tempo (e.g., Banse & Scherer, 1996; Breitenstein et al., 2001).

Like with the *happy* stimuli, a value between the *sad* and *neutral* stimuli was calculated for tempo and tested, but oddly the *sad* stimuli were shorter than the neutral between recordings (0.81, meaning a 20% reduction in tempo). It was hypothesized that this was because of the time limit imposed at the time of the recording session, which caused the speaker to produce utterances at a faster speech rate than would be naturally expected, as this turn of events goes directly against what is found in the literature. Reversing the ratio and slowing the speech down by 20% was also deemed too dramatic when applied across the whole utterance. *Sad* therefore had a 10% increase in duration (i.e., the utterance was made 10% slower), based on measures from the literature and, again, on listener feedback and further testing.

It bears mentioning that tempo differences are not global, but, rather, also sensitive to the prosodic structure of sentences – i.e., it varies depending on its location between accented and non-accented parts of the sentence. Tempo changes were not included in the scope of the present research from the start due to its focus on voice quality and limitations of the software, and will not be explored in depth here. However, it is well known that speech rate is involved in vocal affective signaling, and it was decided to include it to a small, more easily manageable degree in the experiments.

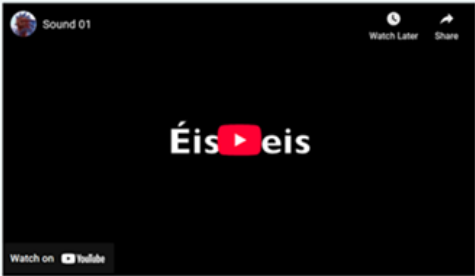
This is also justified by the aim of creating a simple set of changes to be applied to Irish TTS systems, and to avoid further complications in the potential future practical implementation of findings. With these final shifts, the stimuli were then considered ready for testing, and the perception experiment was conducted.

5.3. Perception Tests

The 16 stimuli were presented to 23 listeners, all Irish speakers of varying age, gender and dialect, in a perception test set up on Microsoft Forms. An example of a test question is shown in figure 51.

Triailacha ar Mhothúcháin sa Chaint (Ver.02)

Éist le Fuaim 1 



1 Caidé mar a mhothaíonn an cainteoir? * 

☐ Tá sé ar a sháimhín só (relaxed).

☐ Tá sé brónach (sad).

☒ Tá fearg air (angry).

☐ Tá sé sona sásta (happy).

☐ Tá leadrán air (bored).

☐ Níl aon mhothúchán ar leith i gceist (no emotion).

2 Taispeáin caidé chomh láidir is a bhraitheann sé/sí an mothúchán. * 

1. an-lag	2.	3.	4.	5. an-láidir
☐	☐	☐	☐	☐

Figure 51 - Question example for the discrimination-based perception test.

Listeners were shown a set of instructions and general guidelines in Irish⁷ and then presented with a primer set of stimuli. The primer set consisted of a single file with all stimuli, which listeners were asked to listen to without making any judgements.

This was then followed by six sets, each in which all 16 stimuli were presented in random order. After listening to each stimulus individually, listeners were asked to select the expressed emotion from a list that included all the synthesized affects and a *neutral/no affect* option. If an affect was chosen, listeners were then taken to a follow-up question asking them to rate how strong they deemed the emotion to be on a 5-point scale from weak (1) to strong (5). Questions and multiple-choice options were also presented in Irish. Responses to the first set of stimuli were treated as a practice run to familiarize the listeners with the task and form and discarded for the statistical analysis and results.

5.4. Statistical Analysis and Results

Statistical approaches and results of this perception test are presented below, split into three sections based on each part of the statistical analysis: (1) target affect identification rates, looking into which target affects and voice shift types (global, local and combined) correlated to correct identification of affect by the listener (i.e., which target affect was correctly recognized); (2) perceived intensity rates, examining the effects of target affect and voice shift types on the intensity scores; and (3) a match analysis conducted to examine if intensity scores increased when listeners perceived the ‘correct’ target affect (i.e., when target affect matched listeners’ perceived affect).

The affect identification analysis was conducted using mixed-effects generalized linear logistic regression models (GLMMs) with binomial family distributions and Holm-Bonferroni adjusted significances (Holm, 1979), implemented using the lme4 package in R (Bates, Bolker, & Walker, 2015). The independent variables were target affect and voice shift type. The dependent variable (perceived affect, or target affect identification) was coded as a binary outcome (1 for correct identification, i.e., target affect = perceived affect, and 0 for incorrect identification). The decision for this binary classification was due to issues with rank deficiency, data sparsity and model fitting.

⁷ Translations from English to Irish were done by Prof. Ailbhe Ní Chasaide.

Three separate sets of models were fitted, each looking into different effects in the analysis: (1) overall target emotion performance on affect identification accuracy, investigating unique cueing of each affect and using target affect as a fixed effect; (2) overall voice shift type performance, examining the effects of each shift type on affect identification with shift type as a fixed effect; and (3) affect-specific voice shift type effects, using separate models for each affect (except neutral) to examine which shift type was the most successful at cueing each of the target affects. Interaction effects were not included due to data conversion issues.

Factor variables were relevelled with the following reference categories in the data: 'neutral' (or 'no emotion') for target affect and perceived affect, and 'none' for voice shift type, which was the category associated with the neutral. The models used treatment contrasts with no intercept (-1), meaning that each coefficient represents the log-odds of 'correct' identification for that specific condition relative to chance. The Holm-Bonferroni correction (Holm, 1979), applied to *p* values, was also used to account for multiple comparisons, which was another reason for the choice of the linear model over the multinomial model as a better fit for this experiment.

Affect identification rate results are shown in figures 52-54, and statistical analysis values in tables 8-10. Figure 52 consists of a heatmap confusion matrix showing listener's judgements of affect (perceived affect) based on each target affect and voice shift type. Figure 53 is an alternative representation of the same data, also separated between each target affect for the triad of voice shifts: global (green line), local (orange line), and combined (purple line), with the dotted black line for neutral. These two figures show both the percentages of positive responses obtained for the targeted affect as well as responses for other possible affect choices. The neutral is also shown with the percentages of responses it yielded as each of the possible affect choices. Figure 54 shows the confidence intervals for the target affects and the voice shift types analyses.

PERCEIVED AFFECT RATINGS (TARGET AFFECT + VOICE SHIFT TYPE)

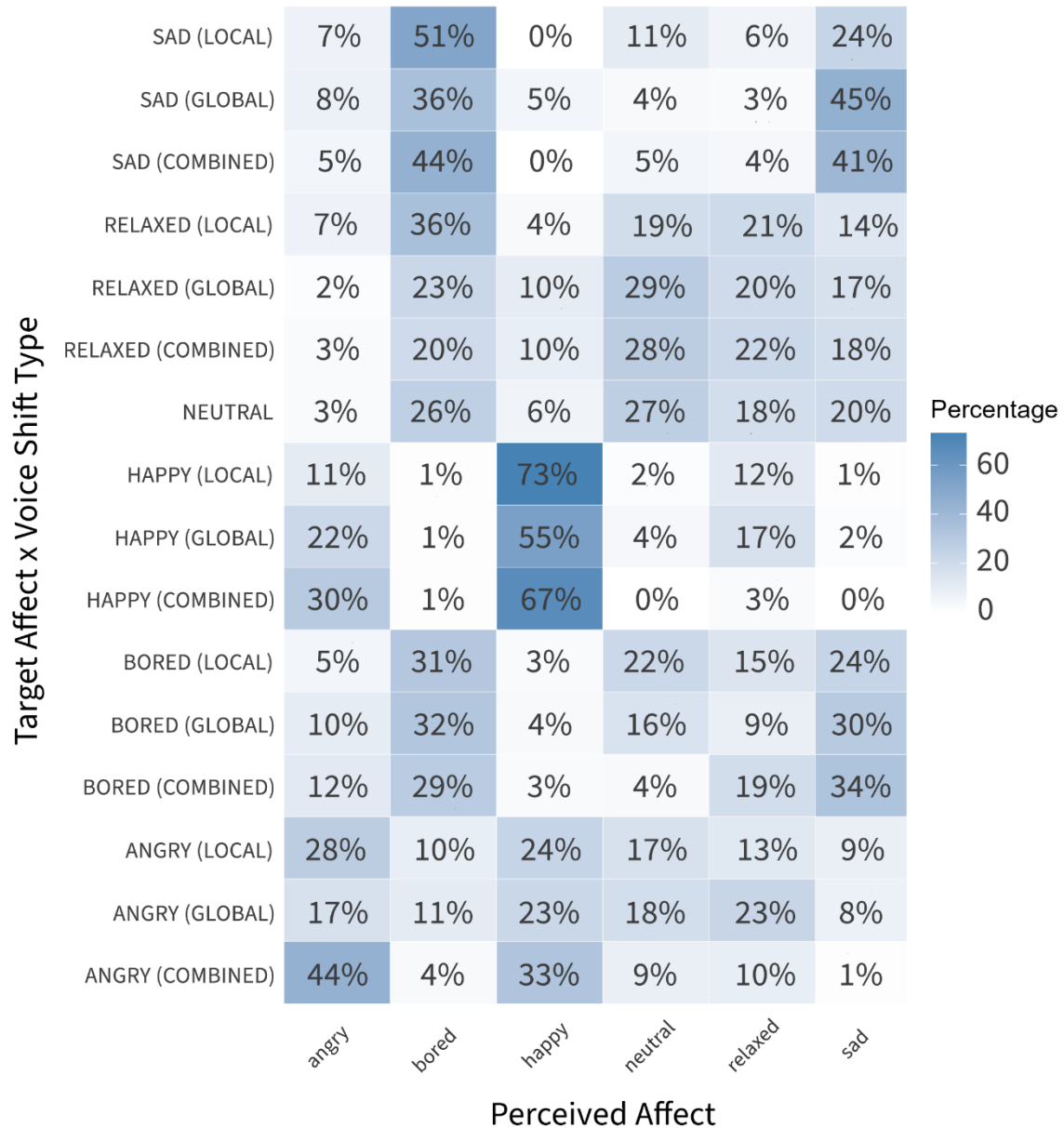


Figure 52 – Affect identification heatmap confusion matrix showing percentages of affect identification rates (perceived affect) based on target affect and shift types.

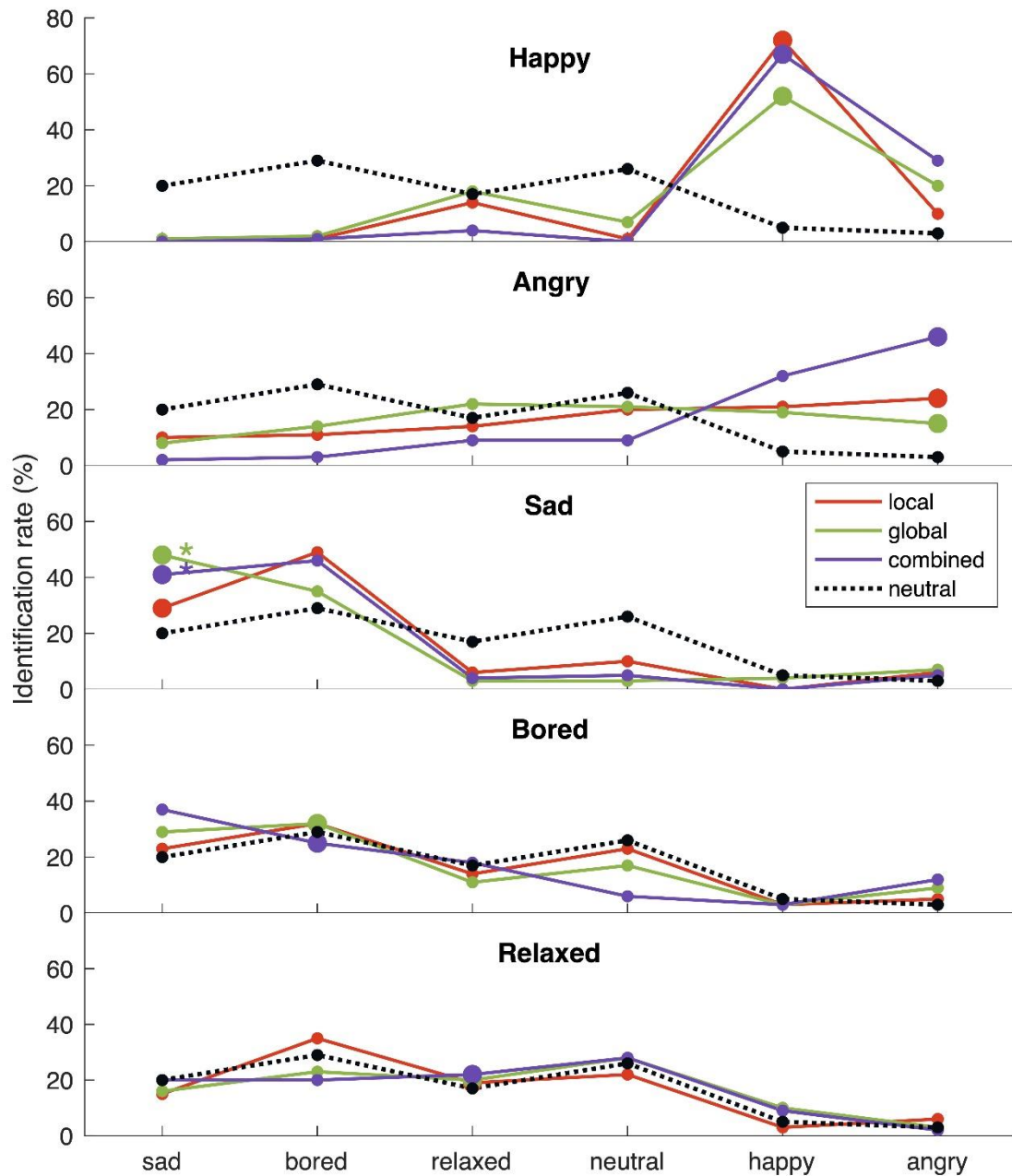


Figure 53 – Alternative representation of affect identification rates for each triad of affect-targeting stimuli. Titles above each section of the image indicate the stimuli's target affect.

Affect listings on the horizontal axis represent perceived affect from listeners, and the vertical axis represents percentages of identification rates. Neutral is shown as the dotted black line for comparison. The larger dots show the values of the target affect for each panel.

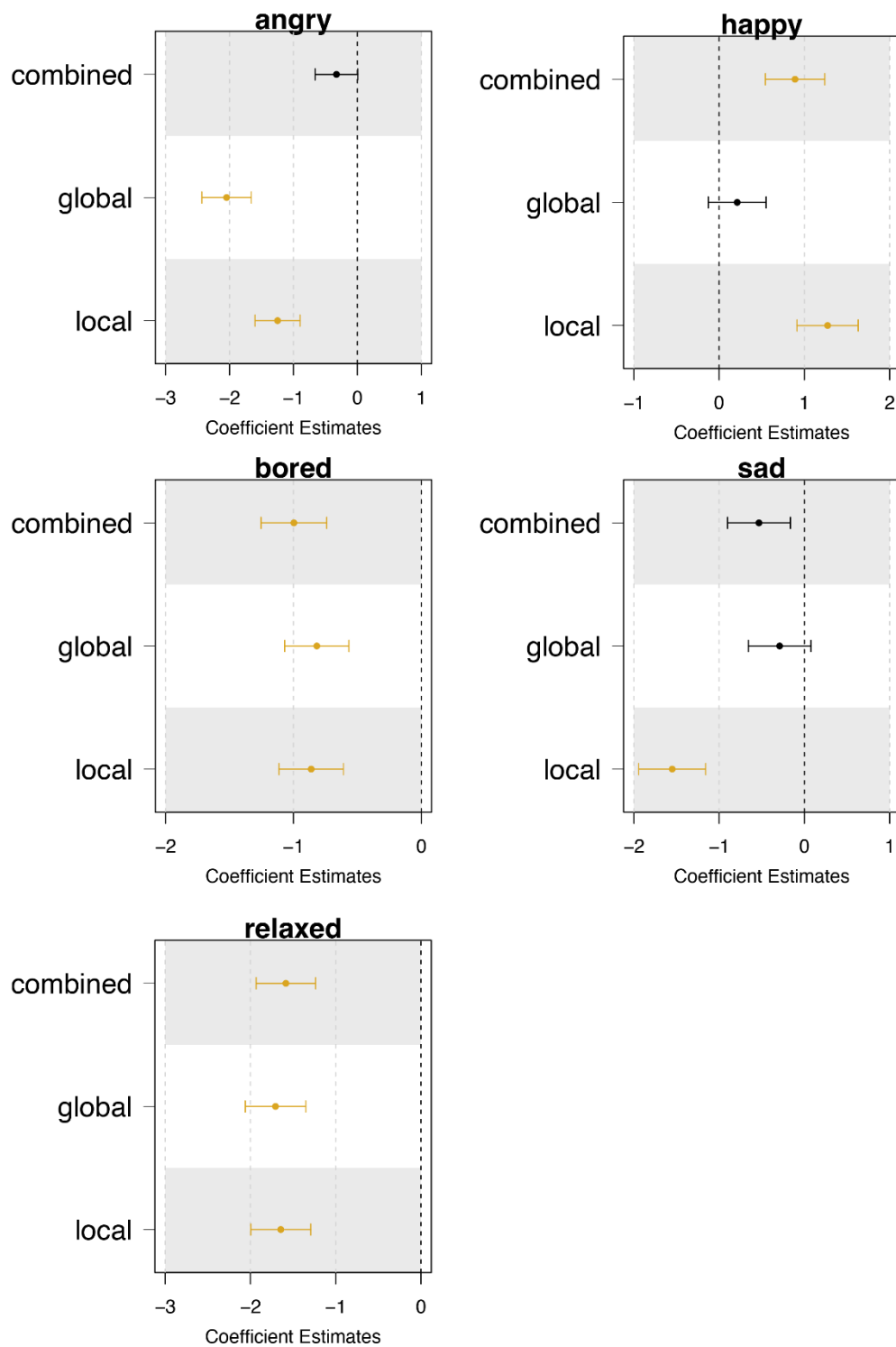


Figure 54 – Identification Rate Coefficient estimates for the voice shift type performance per affect analysis, based on the mixed-effects linear logistic regression analysis conducted for affect identification rates. Yellow lines indicate Holm-adjusted significances. Interaction effects (estimates) are shown on the bottom. Sidebars indicate Standard Error to each side. Bars completely above 0 on the horizontal axis indicates positive correlation, bars intersecting 0 indicate no correlation, and bars below 0 indicate negative correlation (see table 10 for numerical estimate values).

Table 8 – Statistical results for the mixed-effects logistic regression with Holm-Bonferroni adjustments for the target affect identification investigation. Positive estimates indicate above-chance target affect identification performance, indicating unique affect cueing. Negative estimates indicate below-chance target affect identification performance, i.e., no successful unique affect cueing. *** $p < .001$.

Target Affect Performance (Overall Identification Accuracy)						
TARGET AFFECT	Estimate (β)	Std. Error	Z-value	P-value	P-Value (Holm-Adjusted)	Interpretation
<i>Angry</i>	-0.88	0.13	-6.86	<.001	<.001***	Significantly below chance
<i>Happy</i>	0.62	0.12	5.05	<.001	<.001***	Significantly above chance
<i>Bored</i>	-0.82	0.13	-6.48	<.001	<.001***	Significantly below chance
<i>Sad</i>	-0.55	0.12	-4.48	<.001	<.001***	Significantly below chance
<i>Relaxed</i>	-1.35	0.14	-9.51	<.001	<.001***	Significantly below chance
<i>Neutral</i>	-1.01	0.22	-4.66	<.001	<.001***	Significantly below chance

Table 9 – Statistical results for the mixed-effects logistic regression with Holm-Bonferroni adjustments for the voice shift type performance investigation. Negative estimates indicate below-chance target affect identification performance, i.e., no successful unique affect cueing. *** $p < .001$. All voice shift types performed significantly below chance, with *combined* showing the best performance. ‘None’ only applies to the neutral stimulus.

Voice Shift Type Performance (Overall Identification Accuracy)						
TARGET AFFECT	Estimate (β)	Std. Error	Z-value	P-value	P-Value (Holm-Adjusted)	Interpretation
<i>Combined</i>	-0.39	0.10	-4.09	<.001	<.001***	Significantly below chance
<i>Global</i>	-0.68	0.10	-6.97	<.001	<.001***	Significantly below chance
<i>Local</i>	-0.60	0.10	-6.24	<.001	<.001***	Significantly below chance
<i>None</i>	-1.01	0.22	-4.68	<.001	<.001***	Significantly below chance

Table 10 - Statistical results for the mixed-effects logistic regression with Holm-Bonferroni adjustments for the voice shift type performance by target affect. Each coefficient tests whether identification performance with that voice shift type is significantly different from chance level (neither above nor below chance). Positive estimates indicate above-chance performance; negative estimates indicate below-chance performance. Neutral was excluded as it was just the baseline with no voice shifts applied. * $p < .05$, ** $p < .01$, *** $p < .001$.

<i>Voice Shift Type Performance by Target Affect</i>					
TARGET AFFECT + SHIFT TYPE	<i>Estimate (β)</i>	<i>Std. Error</i>	<i>Z-value</i>	<i>P-value</i>	<i>Interpretation</i>
ANGRY					
<i>Combined</i>	-0.33	0.33	-0.98	.33	Not significantly different
<i>Global</i>	-2.05	0.39	-5.29	<.001***	Significantly below chance
<i>Local</i>	-1.25	0.35	-3.55	<.001***	Significantly below chance
HAPPY					
<i>Combined</i>	0.89	0.35	2.56	.01*	Significantly above chance
<i>Global</i>	0.21	0.34	0.63	.53	Not significantly different
<i>Local</i>	1.27	0.36	3.54	<.001***	Significantly above chance
BORED					
<i>Combined</i>	-0.10	0.26	-3.88	<.001***	Significantly below chance
<i>Global</i>	-0.82	0.25	-3.26	<.001**	Significantly below chance
<i>Local</i>	-0.86	0.25	-3.41	<.001***	Significantly below chance
SAD					
<i>Combined</i>	-0.53	0.37	-1.44	.15	Not significantly different
<i>Global</i>	-0.29	0.37	-0.79	.43	Not significantly different
<i>Local</i>	-1.55	0.39	-3.95	<.001***	Significantly below chance
RELAXED					
<i>Combined</i>	-1.59	0.35	-4.55	<.001***	Significantly below chance
<i>Global</i>	-1.71	0.35	-4.81	<.001***	Significantly below chance
<i>Local</i>	-1.65	0.35	-4.68	<.001***	Significantly below chance

For the intensity analysis, a mixed-effects ordinal regression model was employed, using the ordinal package (Christensen, 2023) in R. The intensity ratings were treated as an ordered categorical variable with levels between 0-5 (0 only for neutral), and with an intensity of 50% as the baseline level. Like the identification rate analysis, this model was also split into sections: one to analyze the effects of target affect (*angry, bored, happy, relaxed* and *sad* vs. the neutral reference) on intensity ratings, another investigating voice shift type as fixed effects (with combined as the reference level), and their interaction. Participant ID was included as a random intercept to account for individual differences in rating patterns. Due to design matrix rank deficiency, some coefficients could not be estimated, but testable contrasts were evaluated with Holm-Bonferroni correction for multiple comparisons ($\alpha = .05$). Confidence intervals for both variables and interactions are illustrated in figure 55. Statistical results can be found in table 11.

Table 11 – Statistical results for the mixed-effects ordinal regression analysis with Holm-Bonferroni adjustments, showing main effects of fixed variables and interactions on listeners’ perception of affect intensity. Reference categories are *neutral* for the affect analysis, and *combined* for the voice shift type. Interactions between *sad* and voice shift types were dropped due to rank deficiency, as it showed patterns too similar to the reference levels (i.e., would not have been significant). Significant values = $p < .05$.

Target Emotion × Voice Shift Type Effects on Intensity Ratings		
Predictor	P-value	Significant?
Main Effects		
Target Affect: Angry	.02	✓
Target Affect: Happy	<.001	✓
Target Affect: Bored	.01	✓
Target Affect: Sad	<.001	✓
Target Affect: Relaxed	.58	
Voice Shift: Global	1.00	
Voice Shift: Local	.06	
Interactions		
Angry vs. Global	1.00	
Bored vs. Global	1.00	
Happy vs. Global	.01	✓
Relaxed vs. Global	1.00	
Angry vs. Local	1.00	
Bored vs. Local	1.00	
Happy vs. Local	1.00	
Relaxed vs. Local	.001	✓

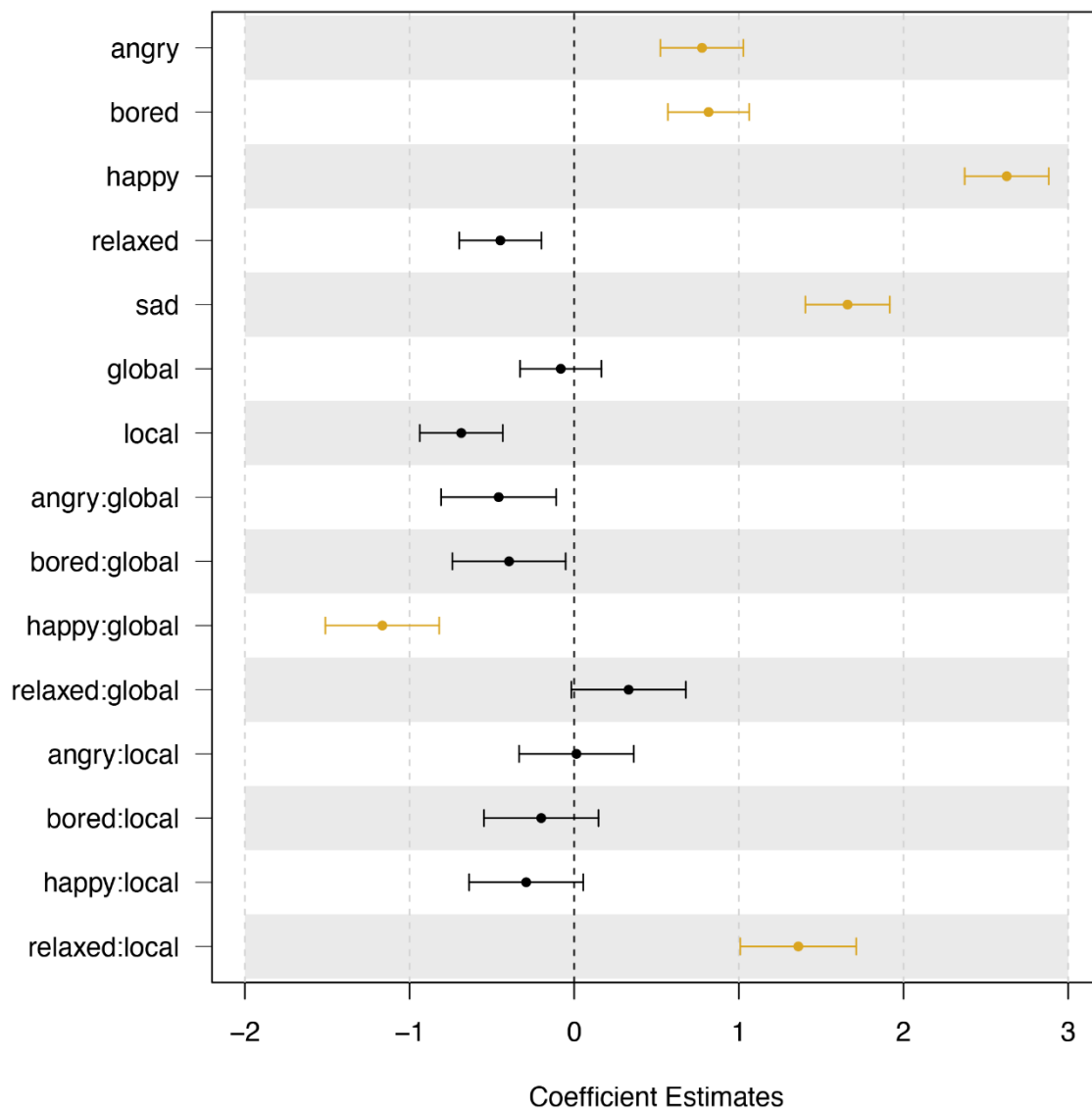


Figure 55 – Intensity Coefficient estimates for target affect, voice shift type and interaction effects, based on the mixed-effects logistic regression analysis conducted for affect intensity rates. Yellow lines indicate Holm-adjusted significances. Sidebars indicate Standard Error to each side. Bars completely above 0 on the horizontal axis indicate positive correlation, intersection with 0 indicate no correlation, and values below 0 indicate negative correlation (see table 11).

Finally, a match analysis using a similar mixed-effects ordinal regression model was conducted to investigate the effects of ‘correct’ affect identification on intensity scores (i.e., if intensity scores were significantly higher when listeners’ perceived affect matched the target affect of the stimuli. This model included correct emotion identification (coded as 1 for matches between target and perceived emotion, 0 for mismatches) as a fixed effect, with participant ID as a random intercept (repeated measures). Model fit was assessed

using AIC and log-likelihood values, and random effects variance was examined to quantify between-participant variability. Results are shown in table 12 below, and in figure 56 in the form a bar plot showing intensity score distributions between correct and incorrect target affect identification.

Table 12 - Statistical results for the mixed-effects ordinal regression analysis with Holm-Bonferroni adjustments, showing the positive effect of correct affect identification (target affect = perceived affect) on intensity ratings. Significant values = $p < .05$.

<i>Correct Affect Identification Effects on Intensity Ratings</i>					
<i>Predictor</i>	<i>Estimate (β)</i>	<i>SE</i>	<i>Z-value</i>	<i>P-value</i>	<i>Significant?</i>
Correct Affect Perception	0.73	0.09	8.03	<.001	✓

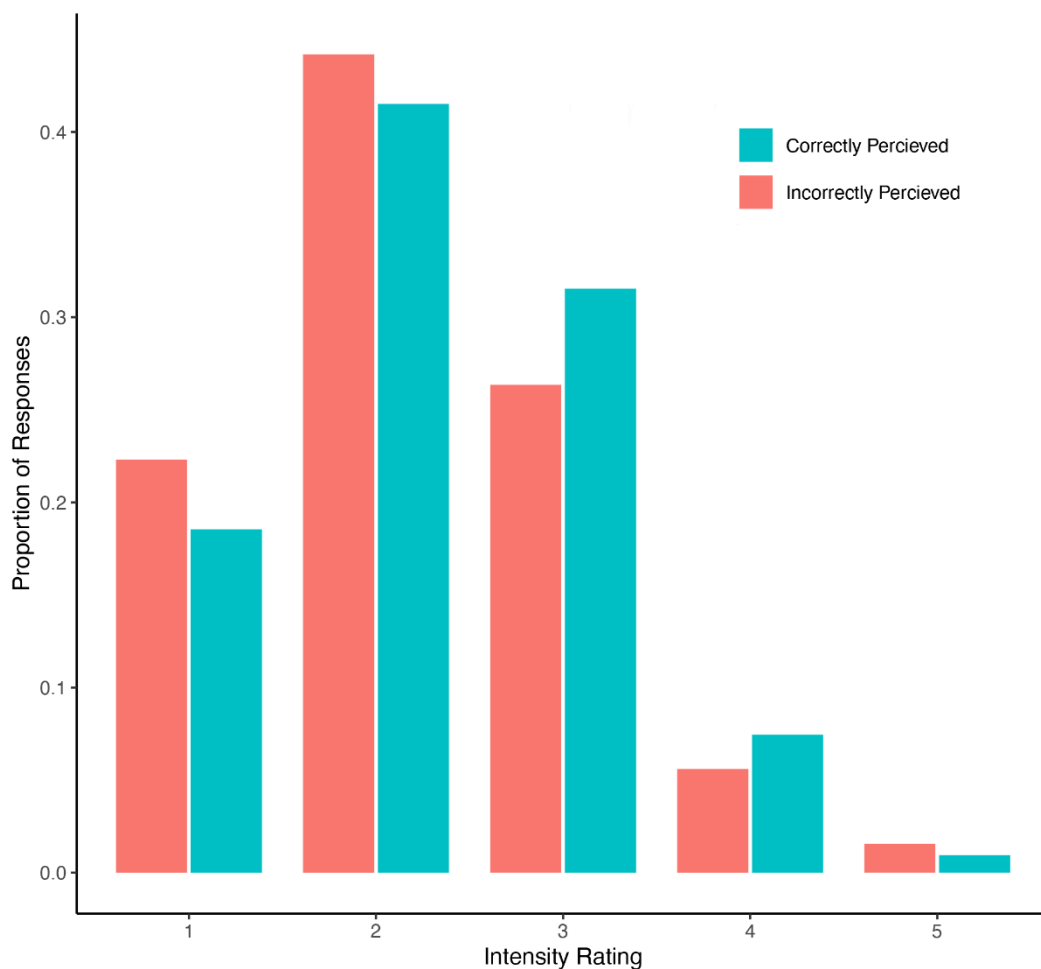


Figure 56 – Intensity score distribution (proportion of responses) of correct vs. incorrect target affect identification.

With these in hand, an interpretation of the results is presented, starting with a discussion on affect identification rates. From the overall analysis, combined voice shifts performed best among the voice shift types ($\beta = -0.39$, not accounting for individual affects), followed by the local and the global ($\beta = -0.60$ and -0.68 , respectively) according to values in table 9. In terms of target affects, the high activation states (*happy* and *angry*) were more successfully cued than the low activation states (*sad*, *bored* and *relaxed*).

Based on the overall target affect analysis (table 8), *happy* was the only affect that showed significant above-chance recognition ($\beta = 0.62$, $p < .001$). As seen in figures 52-53, the *happy* stimuli were accurately recognized 73% of the time for the local stimulus and 67% for the combined, yielding the highest ratings among all the affects. These stimuli were the only ones to achieve significant target affect cueing based on the per-affect analysis (table 10), with estimates of 1.27 ($p < .001$) and 0.89 ($p < .05$) respectively. They therefore appear to cue happiness unambiguously – i.e., distinctly from the other high activation affect, *angry*. The local *happy* stimulus, the most successful at cueing the affect, was only perceived as *angry* 11% of the time, and as *relaxed* 12% of the time. For the combined *happy*, this difference is less striking, but still significant. On the other hand, despite its high recognition rate shown in figures 52-53 (55%), the global *happy* was not uniquely cued to a significant degree ($p = .53$).

Due to the additional modifications to tempo and formant frequencies, it is hard to say how successful the voice source changes would be on their own. Nevertheless, such high ratings for *happy* were surprising, as this affect has been notoriously elusive to capture and cue in perception tests. It was initially hypothesized that the combined stimuli would be the most successful at conveying the target affects. For *happy*, in this experiment, this hypothesis was not supported, as the local modification was the most successful in terms of identification rate and uniqueness of affect signaling.

Despite showing a high recognition rate for the combined stimulus (44%), the *angry* affect was not as effectively signaled as *happy*, nor was it accurately perceived to a significant degree ($\beta = -0.88$, significantly below chance from the overall analysis in table 8, with ‘below chance’ meaning that the affect was ‘correctly’ identified at rates worse than chance relative to the neutral reference across overall shift types). Taking only the recognition percentages into consideration, the initial hypothesis holds – the combined

angry stimulus was the most effective, followed by the local and the global (28% and 17% respectively). This is reflected in the per-emotion analysis estimates (table 10), where combined, with an estimate of -0.33 ($p = .33$), was not significant, but did not harm affect recognition as the local and global shifts appear to have done ($\beta = -1.25$ and -2.05 , respectively). For *angry*, it does seem that the global and local changes were additive, given their poor individual ratings but the relative success of the combined stimulus. This is in line with findings from a previous in-house experiment (Murphy et al., 2022).

The low activation affects received much lower recognition ratings than the high activation affects and were also more ambiguous within their activation categories. Out of the three, the *sad* stimuli achieved relatively strong signaling, with ratings of 45% for the global *sad* stimulus and 41% for the combined *sad* stimulus. Statistically speaking, *sad* was the second affect most effectively cued overall ($\beta = -0.55$), albeit its target affect signaling did not achieve statistical significance. Likewise, values in table 10 confirms that the global shift type was the most effective ($\beta = -0.29$), followed by the combined shift ($\beta = -0.53$). Again, the initial hypothesis was not supported, as adding the local shifts actually reduced the recognition rate of the *sad*-targeting stimuli significantly ($\beta = -1.55$, significantly below chance). The cueing of these stimuli was also not affect-specific: the local and combined *sad* stimuli were more frequently identified as *bored* than even the *bored*-targeting stimuli themselves (49% and 46% respectively).

The recognition ratings for the *bored* stimuli were very low, none being much higher than ratings for neutral. The *bored* affect was much more often associated with the local and combined *sad* stimuli than any of the *bored*-targeting stimuli. In fact, the *sad*-targeting local shifts contributed to the perception of the local *sad* as *bored* (51% rating, seen in figure 52) and reduced its recognition as *sad*, while the addition of global changes had the opposite effect, reducing the perception of *bored*. This suggests that those local shifts applied to this baseline are achieving a fairly successful cueing of boredom and performed much better than the transforms designed to signal this affect. While the overall target affect analysis showed that the *bored*-targeting stimuli performed slightly better than the *angry*-targeting stimuli in cueing the target affect ($\beta = -0.82$, from table 8), all transformations analyzed on a per-affect basis (table 10) showed significantly below chance performance, with estimates very close together.

The stimuli targeting the *relaxed* affect were entirely ineffective, showing the poorest overall identification performance ($\beta = -1.35$) and across all voice shift types. Global and combined *relaxed* stimuli were more often identified as neutral than as their target affect (figures 52-53), and even the local, which achieved the highest identification rate in the triad, did not show statistically significant identification and differentiation, suggesting that this affect category is inherently difficult to identify regardless of the shift types presented here. It is hypothesized that this is due to the additional modification applied to the *relaxed* changes, i.e., the removal of the high f_0 rise in the pre-focal material, which was the main difference between the production stimuli of that affect and the neutral baseline. Without that rise, all parameters were very close to the neutral baseline, therefore resulting in little differentiation from the baseline.

In general, the results for the low-activation states are confounded by the fact that the neutral baseline was generally not perceived as neutral, but about as likely to be perceived as *sad*, *bored* or *relaxed*. This is also reflected in the smaller difference between the low activation targeting affects and the neutral ratings – even if some of the low-activation stimuli achieve high recognition ratings, the shift from the neutral (black-dotted line) is considerably smaller than for the high activation affects. These findings are in retrospect not surprising, given that the acoustic differences between the neutral stimulus and those of the low activation affects are relatively small.

As for the affect intensity analysis, the ordinal regression analysis revealed significant main effects of target emotion on intensity ratings, based on neutral as the reference level (table 11, figure 55): *happy* showed the strongest effect ($p < .001$), followed by *sad* ($p < .001$). Smaller but significant affects were also found for *angry* ($p = .02$) and *bored* ($p = .01$), while *relaxed* did not significantly differ from neutral. In the overall analysis, voice shift types did not show any significant effects, but local shifts approached significance ($p = .06$).

Two interactions between target affect and voice shift type achieved significance: *happy* global ($p = .01$) and *relaxed* local ($p = .001$), suggesting that the effectiveness of different voice shift types on intensity scores varies depending on the target affect. Combined shifts were not presented as the combined category was used as the reference

level for the analysis (i.e., baseline for comparison); therefore, it is important to note that other voice shift categories shown here are interpreted in relation to the combined shift.

It is important to mention, however, that this analysis does not take perceived affect into consideration, i.e., high intensity ratings for the *sad*-targeting stimuli could have been given when *bored* was chosen as the perceived affect category (as seen in the identification rate results above). For this reason, the match ordinal regression analysis was conducted, with results shown in table 12 and illustrated in figure 56. Results indicate that participants did indeed rate stimuli as significantly more intense when the perceived affect matched the target affect compared to when the target affect was misidentified ($\beta = 0.73$, $p < .001$), suggesting a strong relationship between perceptual target accuracy and subjective affect intensity ratings. Based on the previous analyses, however, it must be considered that much of this significance may come from the *happy*-targeting stimuli, considering their high identification rates and significance in the intensity analysis.

5.5. Discussion and Conclusions

The results for this first experiment are mixed. The transforms for the high activation affects (*happy* and *angry*) seem promising, especially considering how challenging the *happy* affect has been to capture in previous research. *Happy* being the most effective transformation both in terms of identification and uniqueness for the local *happy* stimulus was a surprise. However, it must be noted that this particular transform did not involve only voice source modifications, but also tempo and formant shifts, and further experimentation would be required if one was to investigate the contribution of each of these different factors. For the purposes of this project, the same tempo and formant shifts will be maintained for future stimuli and experiments as part of the set of suggested transforms.

The low activation affects were much less effectively cued. Even though there were stimuli that were strongly perceived as *sad* and *bored*, the latter was not in response to the *bored*-targeting stimuli but, rather, to the local *sad* stimulus. It is possible that adding tempo shifts to the *bored* stimuli (following the part of the literature that associated this affect with slower tempo) could have contributed to better ratings for the *bored*-targeting stimuli, but could also have resulted in even less differentiation between the *sad* and the *bored* stimuli. The fact that the neutral stimulus was often perceived as one of such low activation

states makes the results of this experiment somewhat inconclusive, which leads to the necessity to establish a new, true neutral baseline for upcoming experiments.

The initial hypothesis that the combined stimuli would be the most effective for all affects by adding the effects of both global and local transforms was not supported, except for the *angry*-targeting stimuli. The local stimulus was the most effective for *happy*; the global stimulus was the most effective for *sad*; and, for *bored*, the local *sad* stimulus yielded the highest ratings despite not initially targeting this affect. The *relaxed* stimuli were ineffective and more often identified as *bored* or *neutral*, and no stimulus showed a significant cueing of *relaxed*. At this point, there is no definitive answer as to which type of transforms, if one, will be the most effective, even within activation categories, and one can only speculate that different affects will be more effectively cued by different types of transforms. The same can be said for perception of affect intensity: differences seen in the main effects and interactions in the intensity analysis also suggest that the effectiveness of different voice transforms on perceived intensity varies depending on the target affect.

As mentioned, the transforms used in this research were intentionally kept simple and limited to the three voice source parameters (plus the global tempo and formant shifts mentioned) for potential future application to synthetic voices in assistive devices. Because of this, many other voice quality changes that have been associated with these affects have been excluded (e.g., creaky and whispery voice often found in the signaling of low activation states as found in Gobl and Ní Chasaide (2003b) and Yanushevskaya et al. (2018)). Exploring these additional adjustments, while beyond the scope of this project, could yield further insights into affective and linguistic prosody and are worth pursuing in the future.

Based on these results, the next reported experiment follows the same principles as the one presented above, but utilizes a TTS system output (the same speaker's voice but produced through the ABAIR synthesizer) as the neutral baseline. The ABAIR output was considered much more representative of a true neutral, non-emotional utterance based on informal listening tests, and therefore the same hypothesis will be tested with some adjustments to the stimuli and test format.

To conclude, the experiment described in this chapter and conducted as part of the present thesis was reported in the following publication:

Giovannini, A. M., Wang, Z., O'Reilly, M., Ní Chasaide, A., Gobl, C. (2024) 'Voice Transforms for Affect Control in Irish Speech Synthesis', in *Proceedings of Speech Prosody 2024*. Leiden, the Netherlands, pp. 299-303. doi: 10.21437/SpeechProsody.2024-61.

6. PERCEPTION EXPERIMENT 2 – ABAIR BASELINE

The second experiment followed the base methodology of the previous experiment but utilized the ABAIR TTS synthesizer output as a neutral baseline for the affective modifications. There were two reasons for this. Firstly, as discussed in the preceding chapter, the neutral baseline stimulus of experiment 1 was problematic for yielding ratings that tended towards *sad*, *bored* and *relaxed*, and it was felt that the synthetic output of the TTS voice for this speaker, being based on acoustic models built from a large corpus, should be closer to the speaker’s neutral voice. Secondly, as a practical downstream goal of this research is the formulation of relatively simple transforms that might be used to provide affective coloring of ABAIR’s TTS voices, it made sense to test the affective transforms relative to the baseline voice of the individual TTS voice.

Utilizing the ABAIR TTS synthesizer output as a baseline for the affective modifications required some adjustments to the synthesis process. Additionally, a new format for the perception test was used, described in the following sections.

6.1. The ABAIR Baseline Neutral and Modification Details

Because a new baseline was being utilized for this experiment, a few adjustments were needed before applying the affect transformations. First, a new manual analysis of the neutral baseline was required, as the VSG system (Z. Wang & Gobl, 2023b) requires a set of parameter values as input and not just the audio file. Because the ABAIR synthesizer does not provide voice source parameters through the synthesis process itself, the downloaded audio from the synthesis had to be entered into the ISF system for source modelling like the original recording from the previous experiment.

The first analysis challenge arose due to the lack of phase linearity in the synthesizer. Unfortunately, the ABAIR synthesis also does not retain the phase linearity of the original recordings, which makes the manual source modelling rather challenging and it had to be conducted with great care. The resulting parameters were satisfactory enough to resynthesize the ABAIR signal through the VSG system successfully, retaining good sound quality and the neutral quality of the ABAIR utterance.

One issue that resulted from the lack of phase linearity and was identified after resynthesizing the neutral through the VSG, however, were very low EE values and consequently high RD values. This caused the volume of the resynthesis through the VSG to be lower than ideal, and to have a considerably laxer voice quality. To correct for this, it was decided to rescale the EE and RD values of the ABAIR neutral. EE values were divided by 0.7 to increase the overall level, while RD values were multiplied by 0.7 to increase the tension in the voice. There was no change to f_0 .

Figures 57-59 below show a comparison between the original recorded neutral, the original ABAIR neutral, and the rescaled ABAIR neutral⁸. There was no rescaling of f_0 done to the original ABAIR synthesis, which is why the two contours overlap below.

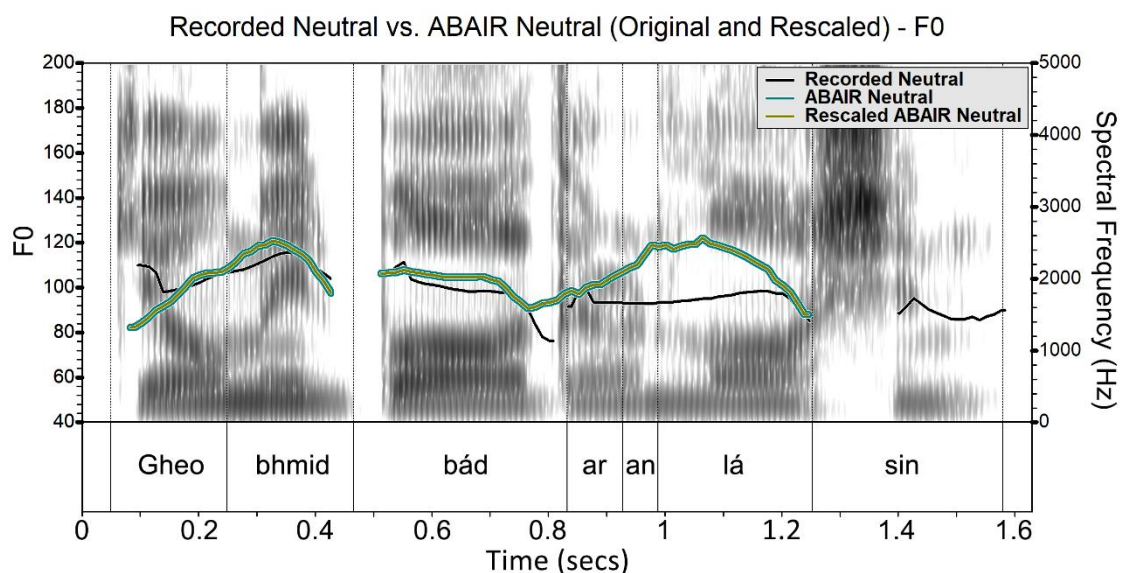


Figure 57 – F_0 comparison between the original recorded Neutral, the ABAIR Neutral, and the rescaled ABAIR Neutral (all resynthesized through the VSG).

The major difference seen between the ABAIR synthesis and the original recording is the very clear peak on the second pitch accent '*lá*' produced by the synthesizer, following the standard intonation patterns of this dialect. In figure 57, illustrating f_0 , it goes even

⁸ Parameter illustrations in figures 57-59 were created through the custom Praat script by Rodgers (2022) mentioned previously and edited through Adobe Photoshop (Adobe Inc., 2023) for additional formatting adjustments.

higher than the first accent '*bád*'. The rest of the sentence has slightly higher peaks, but overall, not as dramatic as the difference seen in '*lá*'.

For EE (figure 58), the non-rescaled ABAIR neutral utterance is similar in contour to the recorded stimulus at the starting, pre-nuclear section of the utterance. For the accented syllables, there seems to be a small reversal – the accent on '*bád*' is flatter in the ABAIR neutral, with a dip on the first half of the vowel followed by a rise, while the recorded neutral features a steadier fall. On the other hand, the opposite happens on '*lá*' – the recorded neutral shows a subtle but steady rise, while the contour of the ABAIR neutral, albeit with some fluctuations, show a high start followed by a fall.

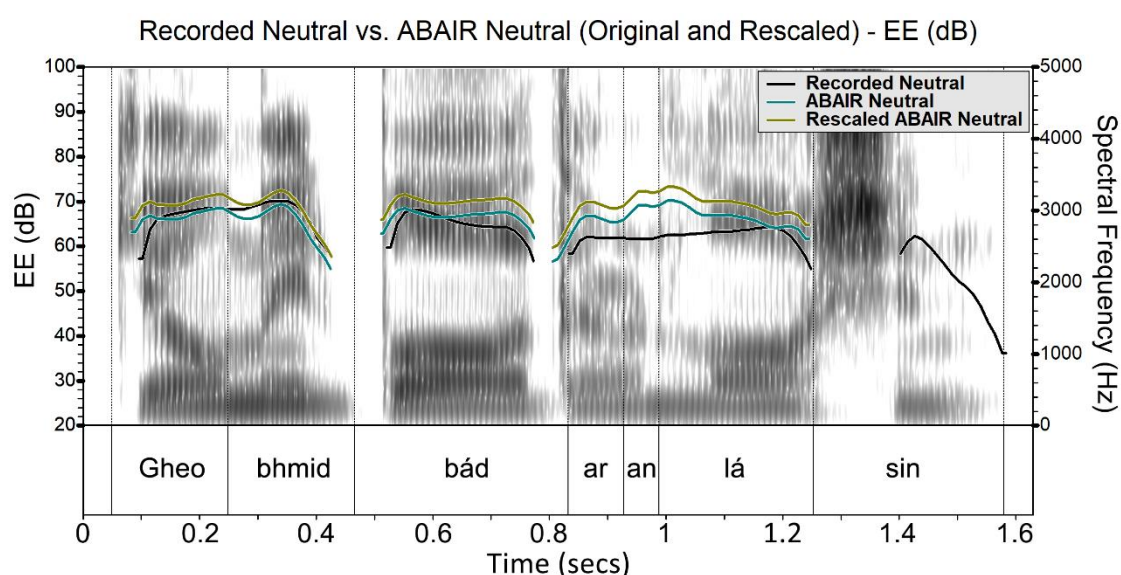


Figure 58 – EE (dB) comparison between the original recorded Neutral, the ABAIR Neutral, and the rescaled ABAIR Neutral (all resynthesized through the VSG).

Finally, figure 59 illustrates RD. The pre-focal material in the ABAIR neutral has a sharp rise, similar to the recorded neutral. Curiously, the first pitch accent on the ABAIR neutral features a rounder contour dip that almost mirrors the one from the recording, which shows a sharp rise and fall around the same part of the vowel. On the other hand, the second accent '*lá*' maintains a slowly falling contour in both stimuli.

Overall, differences in parameter levels are not directly comparable due to the different methods of acquisition of each stimulus. With this, it was confirmed that the

rescaling was working properly through the VSG system, and the rescaled ABAIR resynthesis was deemed a proper baseline for the perception experiment.

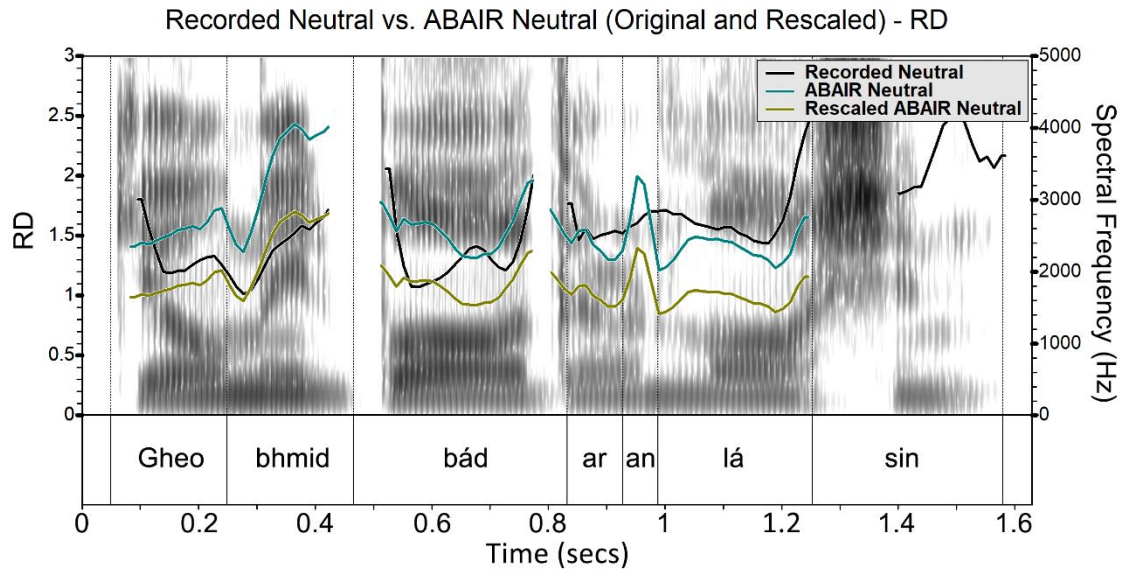


Figure 59 – RD comparison between the original recorded Neutral, the ABAIR Neutral, and the rescaled ABAIR Neutral (all resynthesized through the VSG).

One additional change from the previous experiment was that the f_0 peak of the local shifts for the *angry* stimulus was subtly increased, and RD was slightly lowered at the same time point. Whereas the previous f_0 scaling factor for the first anchor point of the angry pitch accents was 1.3, resulting in a f_0 of approximately 150 Hz, this anchor point was now raised to 1.5 to strengthen the perceived affect, bringing the peak up to approximately 174 Hz and maintaining the subsequent fall. The RD scaling factor was changed from 0.65 at the same anchor point to 0.6, slightly increasing tension in the voice.

Tables 13 and 14 present global and local shifts numerically with the updated neutral baseline and the small adjustments to the scaling factors of *angry*. Figures 60-62 and 63-65 show the global slopes and the local contours respectively, in a similar manner as done in section 5.1 but made more compact given the similarities.

Table 13 - Global shift values (coefficient and intercept) for each affect and parameters (low regression line for f_0 and EE; high regression line for RD).

Affect	Parameters					
	F_0		EE		RD	
	Coefficient	Intercept	Coefficient	Intercept	Coefficient	Intercept
Neutral	10.0	91.6	0.4	63.0	-0.7	2.3
Angry	-24.4	132.1	-6.9	72.8	0.0	1.4
Happy	-21.9	137.4	0.8	64.6	-0.2	1.8
Bored	-28.6	100.3	-4.1	59.5	0.1	1.7
Sad	-10.1	97.6	-7.3	64.7	0.3	1.7
Relaxed	-32.5	121.8	-4.5	63.9	-0.6	2.4

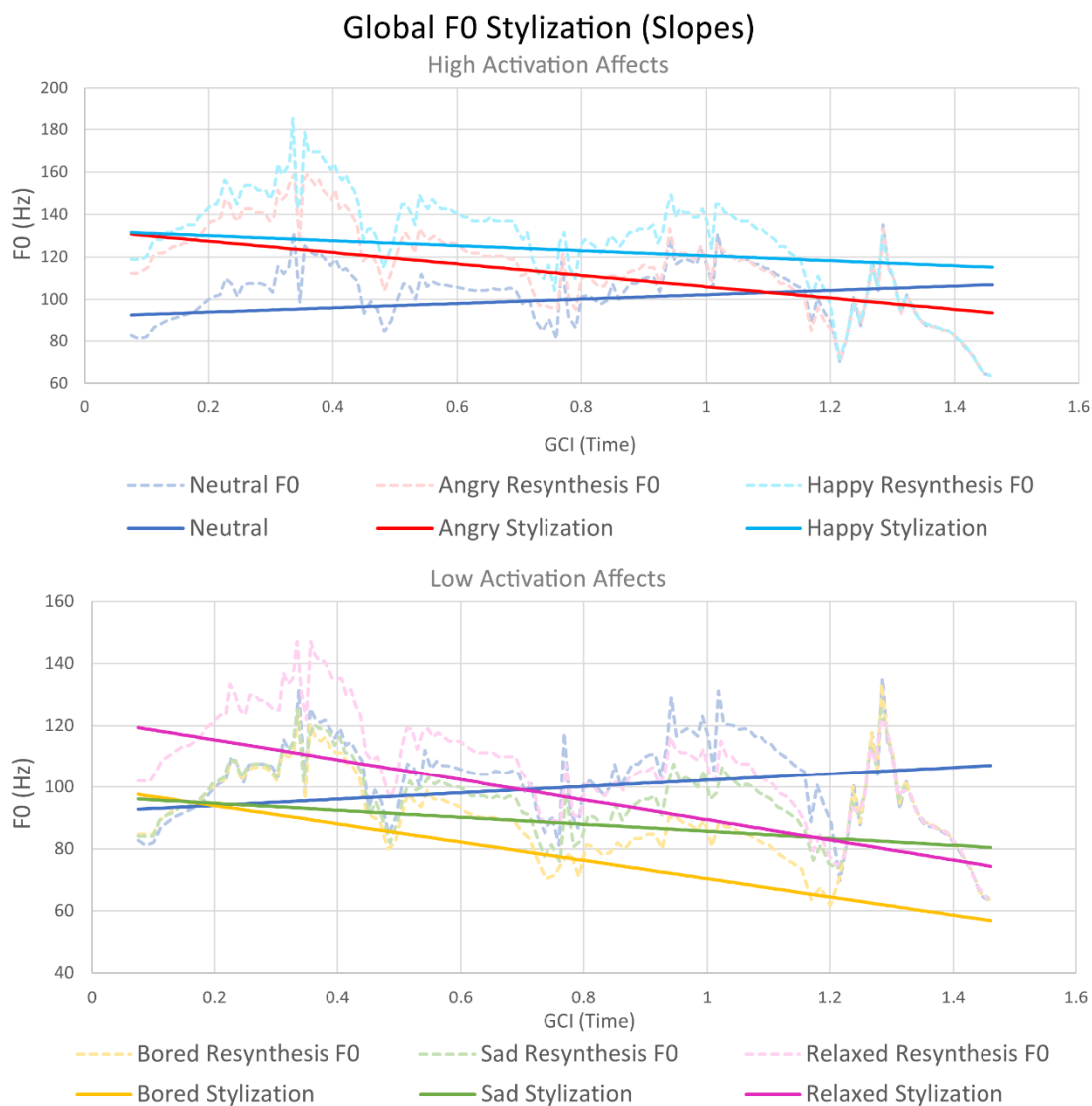


Figure 60 - Global slopes for f_0 , from values in table 10.

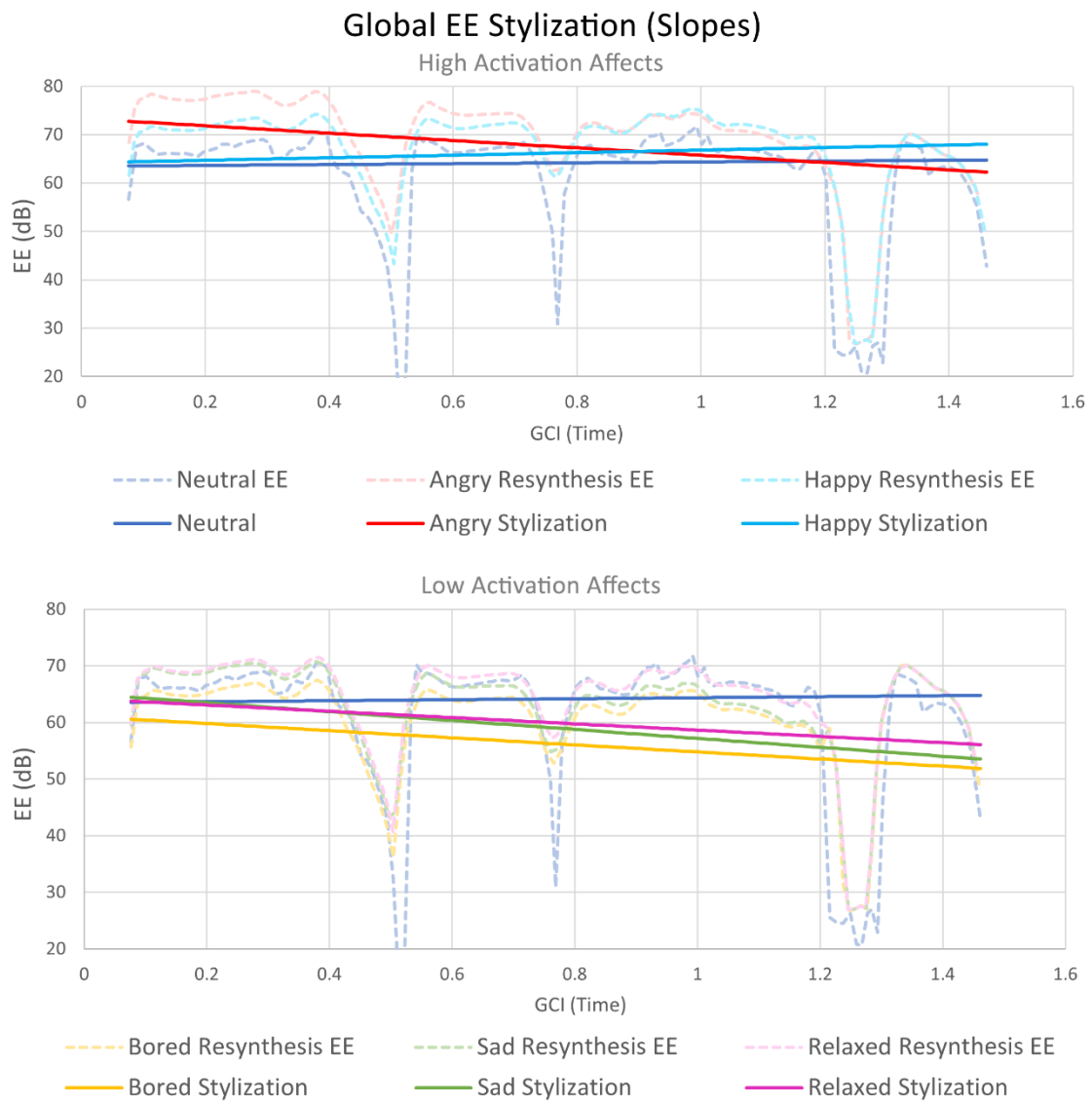


Figure 61 – Global slopes for EE, from values in table 10.

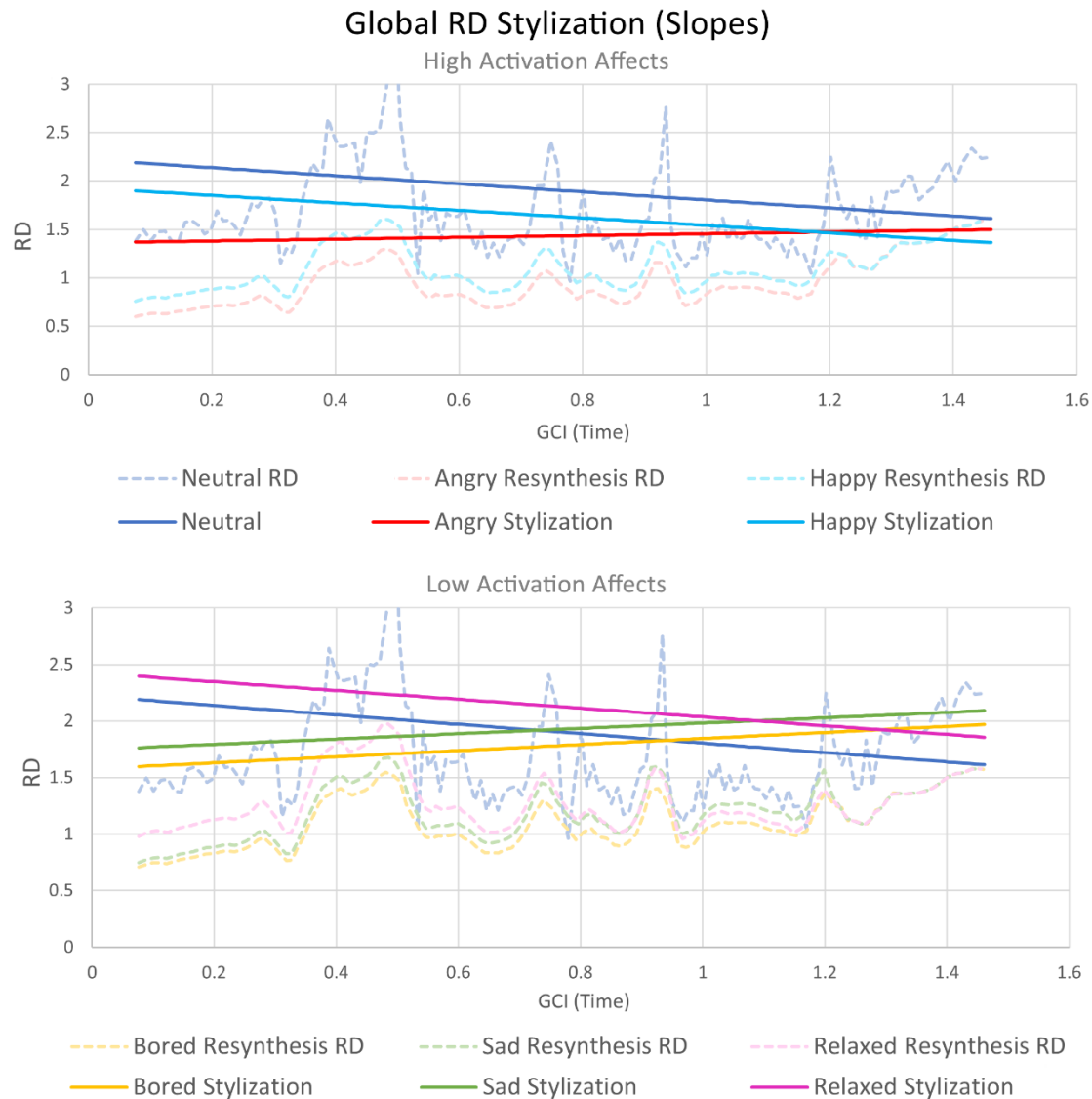


Figure 62 – Global slopes of for RD, from values in table 10.

Global shifts values (table 13) for each of the affects remain the same as those on table 3, chapter 5. The difference here is the ABAIR neutral baseline. F_0 shows a rising regression line for the ABAIR neutral starting from a lower intercept, in contrast to the recorded neutral (which features a coefficient of -13.9, and an intercept of 108.2). For EE, there is little change to the intercept, but the same rise happens for the coefficient (previously at 63.2 and -2.7., respectively). The same inversion is seen in RD – because this parameter is essentially upside down relative to the others, the regression analysis now features a fall, starting from a laxer voice quality (compared to the former coefficient of 0.3 and intercept of 1.5).

Table 14 - Local shift values for each affect, applied as scaling factors over the neutral or target regression line through the anchor point system described in chapter 4. II_1 and II_2 refer to the anchor points 20% into the accented vowels (*bád* and *lá*), from the starting boundary of the vowel and to the ending boundary of the vowel, respectively. III refers to the anchor point in the middle of the vowel, as detailed in section 4.2.

Affect	Parameters & Anchor Points								
	F_0			EE			RD		
	II_1	III	II_2	II_1	III	II_2	II_1	III	II_2
Angry	1.5	1.1	1.0	2.8	2.0	1.5	0.6	0.7	0.8
Happy	1.4	1.5	1.3	2.4	2.4	2.2	0.7	0.8	0.7
Bored	1.1	1.1	1.0	1.5	1.2	0.9	0.8	0.9	1.0
Sad	1.1	1.1	1.0	1.3	1.5	1.5	1.0	0.8	0.8
Relaxed	1.0	1.0	1.0	2.0	1.7	1.25	0.9	0.8	1.0

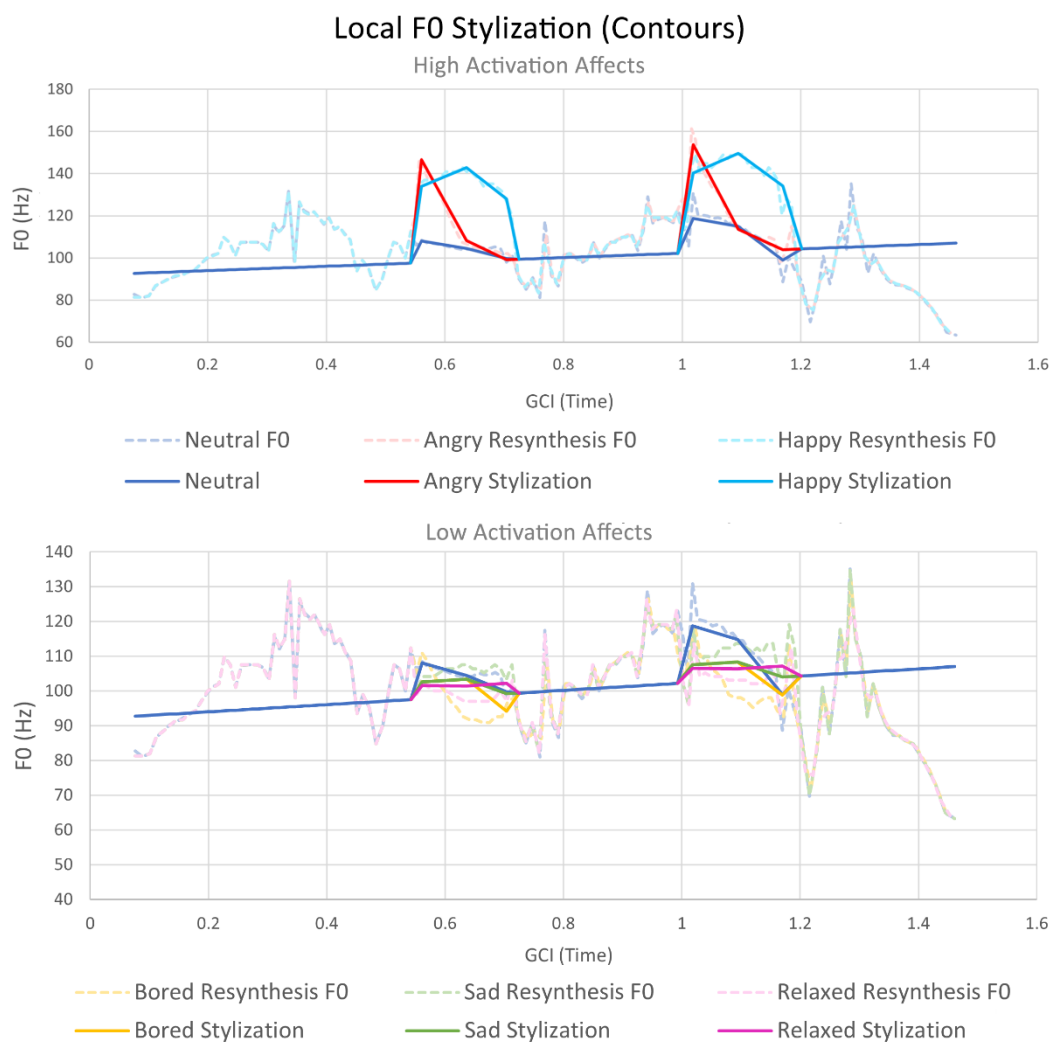


Figure 63 - Local contours for f_0 , from values in table 11.

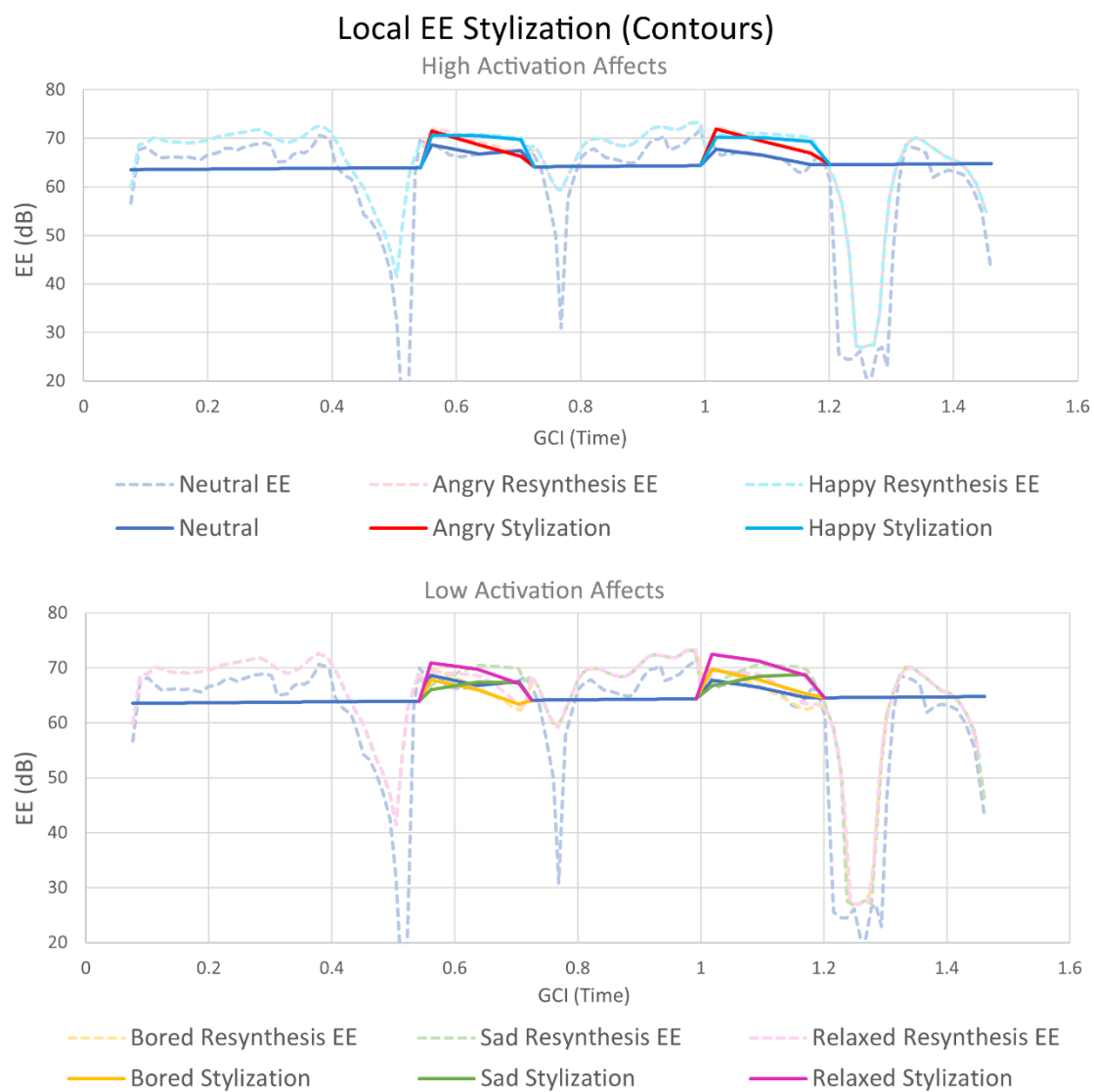


Figure 64 – Local contours for EE, from values in table 11.

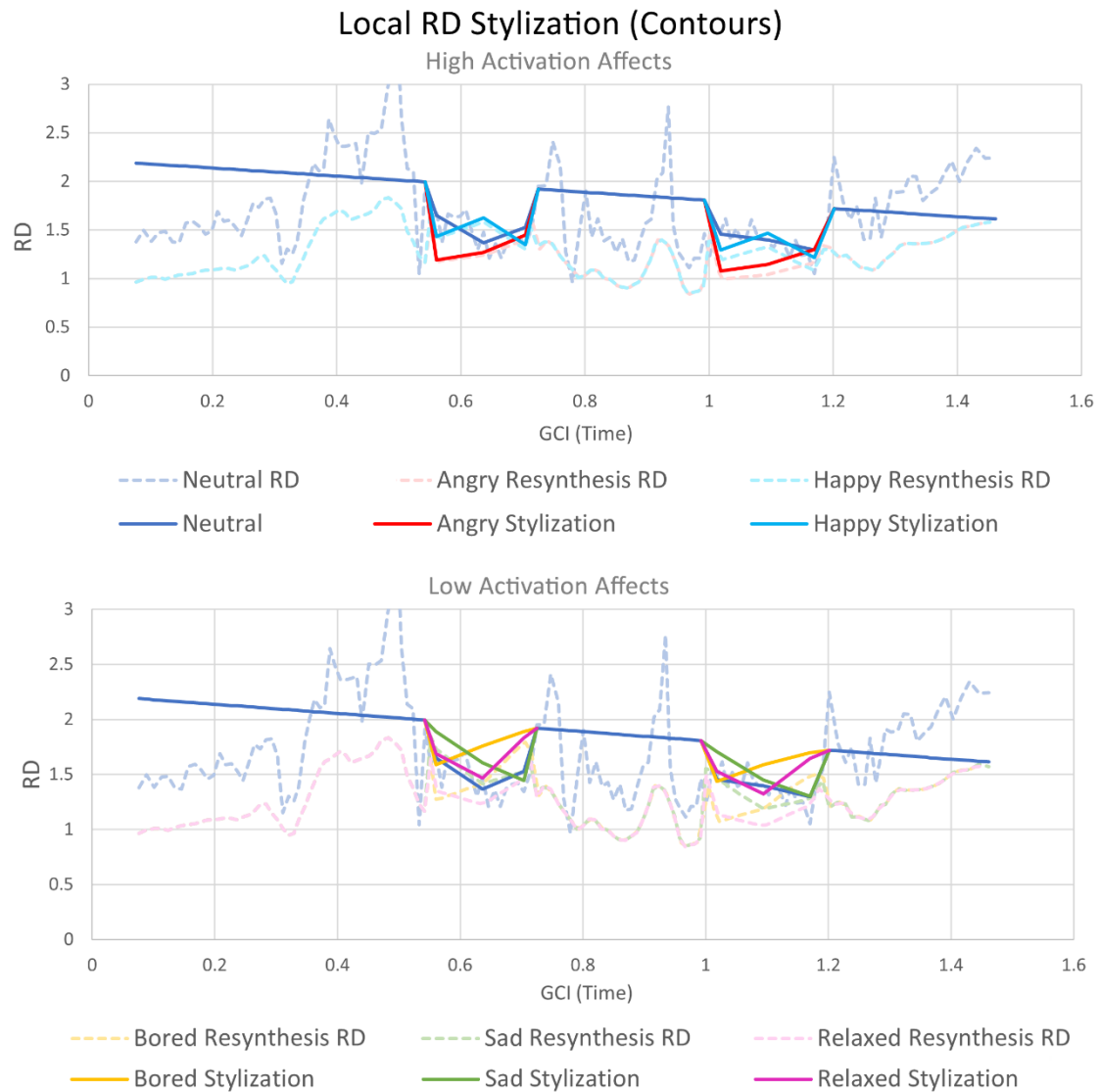


Figure 65 – Local contours for RD, from values in table 11.

Table 14, listing the local shifts, remains largely the same as table 4 in chapter 5, as scaling factors were calculated based on the neutral stimulus lower regression baseline to be applied to the affect-targeting lower regression line. The only differences are those mentioned above – the peak of the *angry* stimulus (f_0 column II_1) was slightly raised, along with a slight lowering of RD at the same anchor point (RD column, II_1). These stylized contours are also illustrated in figures 63-65, with the scaling previously mentioned applied to the resynthesis contours (dotted lines), accounting for the differences between the original neutral and affect stylization contours, and the resynthesis lines from the VSG output.

Table 15 - Mean Difference values taken across the full utterance between the affective stimuli and the neutral baseline. Highlighted values are those most relevant to evaluate the performance of the VSG system in applying the global, utterance-wide shifts.

Affect	Mean Difference Values – Global vs. Local vs. Combined (Full Utterance)								
	F0 (semi-tones)			EE (dB)			RD		
	Global	Local	Combined	Global	Local	Combined	Global	Local	Combined
Angry	2.5	0.3	2.7	4.3	1.0	5.0	0.8	1.0	0.8
Happy	4.2	1.6	5.6	2.1	1.7	3.6	0.9	1.0	0.9
Bored	-2.1	-0.3	-2.3	-5.2	-0.3	-5.6	0.9	1.0	0.9
Sad	-1.1	-0.2	-1.3	-3.0	0.1	-3.3	1.0	1.0	1.0
Relaxed	0.9	-0.2	0.7	-1.3	0.3	-1.3	1.0	1.0	1.1

Table 16 - Mean Difference values taken across the first pitch accent of the utterance *bád* (between anchor points ll_1 and ll_2 defined in table 4), between the affective stimuli and the neutral baseline. Highlighted values are those most relevant to evaluate the performance of the VSG system in applying the desired shifts (i.e, the local shifts applied specifically to the pitch accent *bád* in the local column, and the combined shifts applied to the same pitch accent but with as the addition of the local and global shifts).

Affect	Mean Difference Values – Global vs. Local vs. Combined (First Accent - Bád)								
	F0 (semi-tones)			EE (dB)			RD		
	Global	Local	Combined	Global	Local	Combined	Global	Local	Combined
Angry	2.7	1.0	3.6	4.4	1.9	6.3	0.7	0.9	0.7
Happy	4.9	4.2	9.1	1.8	2.9	4.8	0.9	1.0	0.9
Bored	-2.4	-0.4	-3.1	-6.0	-2.3	-8.3	0.9	1.1	1.0
Sad	-1.2	-0.3	-1.8	-3.5	-1.0	-4.5	1.0	1.1	1.1
Relaxed	1.1	-0.3	0.6	-1.8	0.1	-1.6	1.1	1.1	1.2

Table 17 - Mean Difference values taken across the first pitch accent of the utterance *lá* (between anchor points ll_1 and ll_2 defined in table 4), between the affective stimuli and the neutral baseline, for additional reference.

Affect	Mean Difference Values – Global vs. Local vs. Combined (Second Accent - Lá)								
	F0 (semi-tones)			EE (dB)			RD		
	Global	Local	Combined	Global	Local	Combined	Global	Local	Combined
Angry	-0.1	0.5	0.1	0.6	2.7	3.1	0.9	0.8	0.7
Happy	2.4	3.5	5.8	2.1	3.7	5.8	1.0	0.9	0.9
Bored	-4.9	-1.0	-6.9	-6.7	-1.4	-9.2	1.1	1.0	1.1
Sad	-3.0	-0.9	-3.7	-7.1	-0.1	-7.1	1.3	1.0	1.2
Relaxed	-2.0	-1.0	-2.6	-3.9	0.7	-2.7	1.2	1.0	1.1

Tables 15-17 present the mean difference values between the rescaled ABAIR neutral, and the new affective stimuli produced with the values from tables 13 and 14 in

the VSG system. These values once again seem sensible based on the production data and literature findings, also indicating that the system is performing accordingly. *Angry* and *happy* show increases in f_0 and EE in both global and local terms, albeit to a smaller degree due to the baseline change. They also feature respective decreases for RD, at a higher rate. The lower activation affects *bored* and *sad* show overall decreases in f_0 and EE (now with a bigger difference given the ABAIR neutral baseline), with very slight rises in RD, as would be expected. Finally, *relaxed* is close to the neutral baseline in terms of f_0 and EE, but feature the largest increase in RD, accordingly.

Figures 66-68 illustrate these shifts for the combined manipulation of the *angry* affect, in the form of f_0 , EE and RD contours from the VSG output compared to the new rescaled ABAIR neutral baseline. Illustrations for all affects are found in Appendix D⁹.

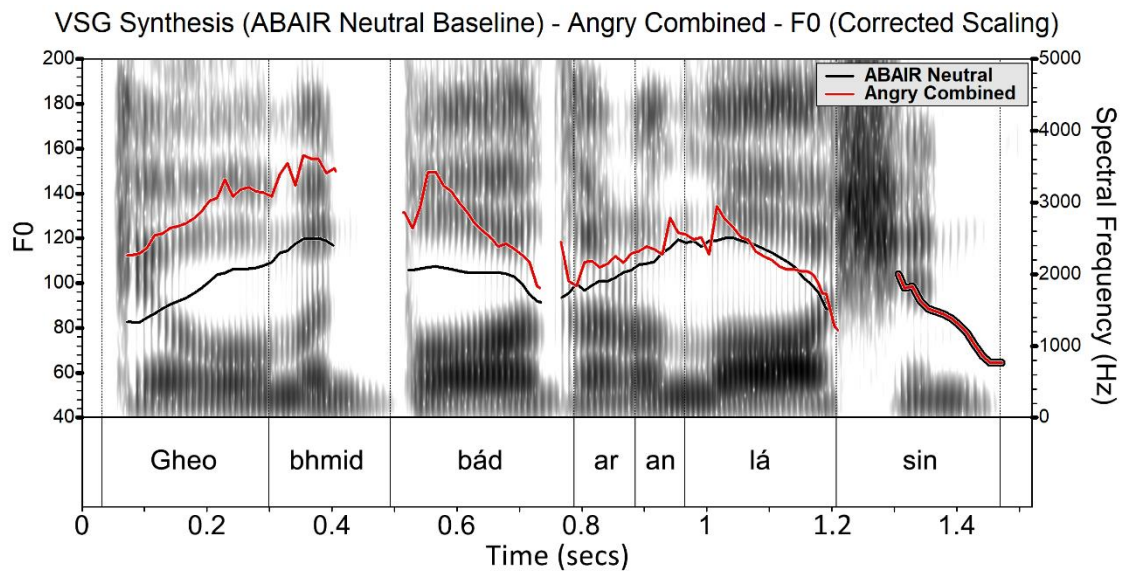


Figure 66 - F_0 contour of the combined modification of the rescaled ABAIR *neutral* stimulus to the *angry* target affect.

⁹ Parameter illustrations in figures 66-68 and in Appendix D were created through the custom Praat script by Rodgers (2022) and edited through Adobe Photoshop (Adobe Inc., 2023) for formatting adjustments.

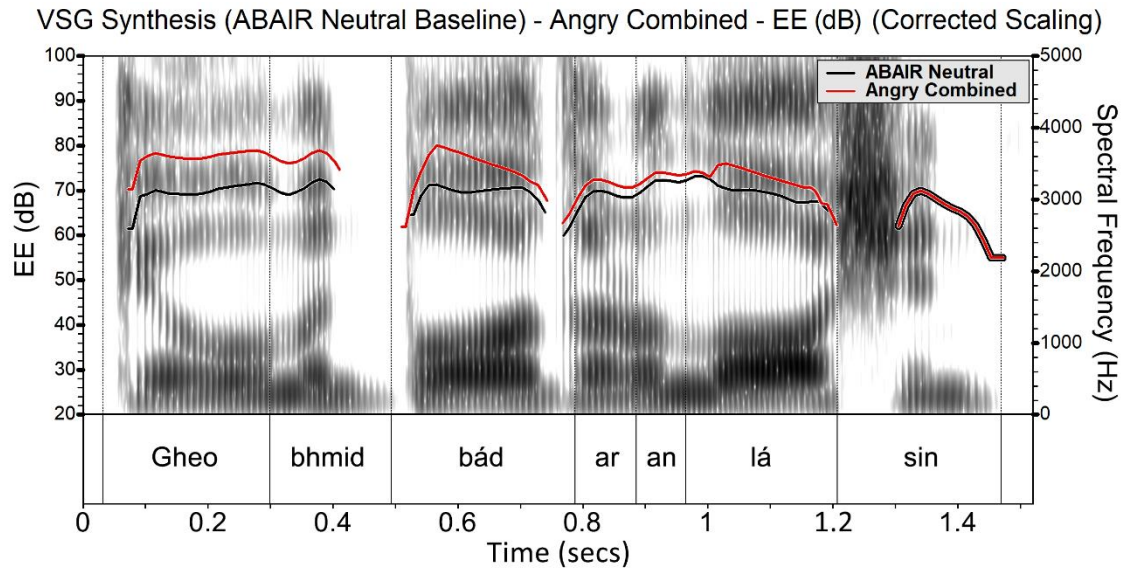


Figure 67 – EE contour of the combined modification of the rescaled ABAIR *neutral* stimulus to the *angry* target affect.

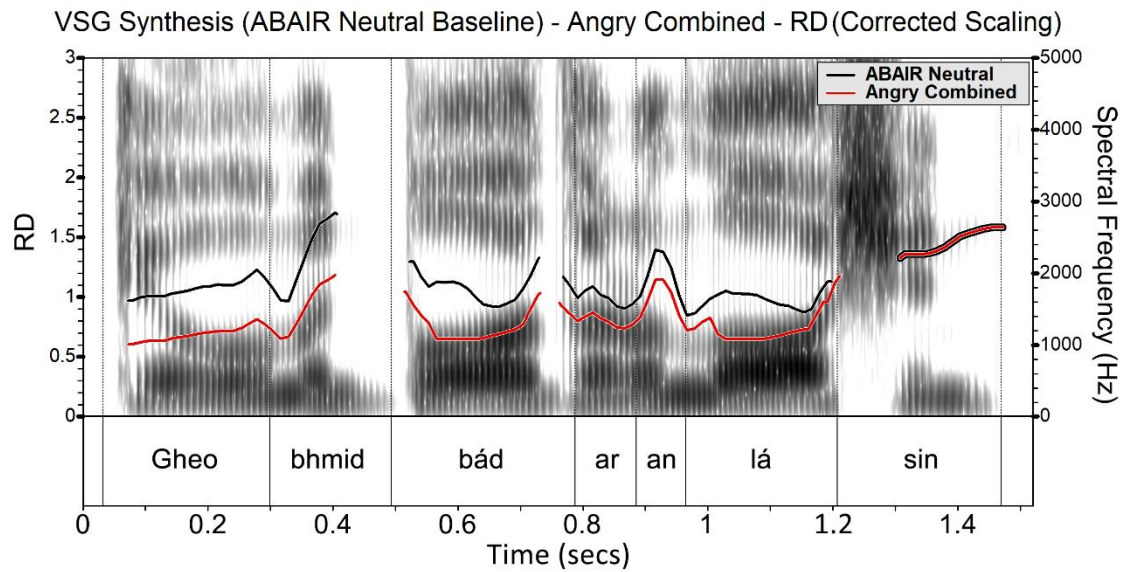


Figure 68 - RD contour of the combined modification of the rescaled ABAIR *neutral* stimulus to the *angry* target affect.

Finally, the same additional formant and tempo shifts for the *happy* affect (4% increase and 10% reduction, respectively) and the tempo shift for the *sad* affect (10% increase) were applied in the same manner as in Experiment 1 in chapter 5.

6.2. Perception Tests

These 16 new stimuli were presented to 34 listener judges with varying levels of proficiency in Irish in a perception test, set up on Google Forms. They were presented with a set of simple instructions in Irish¹⁰, including listening to the stimuli with headphones and in a quiet environment. They were then presented with a primer set of stimuli, consisting of one version of the neutral presented separately, and two sequences of all the presented affects to familiarize themselves with the stimuli, and have the chance to adjust their system volumes based on the lower intensity of the low activation affects. Before starting the test proper, listeners were also presented with a six-question sequence based on the first task as a practice run to familiarize themselves with the task.

This test was divided into three scale-based tasks. Each task included 16 pairs of utterances: the neutral stimulus followed by one of the 16 stimuli (including neutral), presented in a randomized sequence. For each task, listeners were asked to judge the emotion of the second utterance of the pair on a 7-point scale of intensity between -3 (low activation) and +3 (high activation), each polarity representing one of two opposing affects. The scale for task 1 represented the *sad* vs. *happy* affects; task 2 tested the *bored* vs. *interested* affects; and task 3 encompassed the *relaxed* vs. *angry* affects. If listeners perceived one of the two affects in the scale, they were asked to make their judgement on the numerical scale between -3 (a strong expression of the low activation affect) and +3 (a strong expression of the high activation affect). The ± 2 represented a moderate expression of that emotion, and ± 1 represented a slight expression. Listeners were asked to choose the midpoint 0 (zero) whenever no affect was perceived (no emotion/neutral), or when the perceived affect was not on any end of the scale presented in that subtest.

An *interested* option was included as a counterpart to *bored* to allow for a threefold pairing of affects but no stimuli targeting *interested* were included in the set. The inclusion of *interested* as an option was also motivated by the fact that, in the previous experiment, listeners often judged stimuli targeting *angry* as *interested*. Because practice runs were

¹⁰ Translations from English to Irish were done by Prof. Ailbhe Ní Chasaide.

included before each task, no responses to the actual set of 16 stimuli were discarded. An example of a question presented to listeners for task 1 is shown in figure 69.

Ceist 1/16 (Question 1/16)

01

Anna G.

Sa dara abairt, an bhfuil an cainteoir **BRÓNACH** nó **SONA**? agus Cé chomh láidir *
is atá an mothúchán?

In the second sentence, is the speaker SAD or HAPPY? How strong is the emotion?

0 = Níl ceachtar mothchúchán i gceist
0 = Neither emotion

-3 An-Bhrónach
(Very Sad)

-2

-1

0 Níl Ceachtar
(Neither)

+1

+2

+3 An-Sona
(Very Happy)

Roghnaigh
(Choose):

☐

☐

☐

☐

☐

☐

☐

Figure 69 - Question example for the scale-based perception test, showing the scale for task 1 (*happy* vs. *sad*).

184

6.3. Statistical Analysis and Results

Statistical approaches and results of this test are presented below. For the statistical analysis, mixed-effects ordinal logistic regression models were used to measure the relationships between the intensity scores and target affects/voice shift types through R (R Core Team, 2021). Three separate models were fitted, one for each affect dimension present in the perception test: *sad-happy*, *bored-interested*, and *relaxed-angry*. The 7-point scale (between -3 and +3) used for participant ratings was treated as an ordinal dependent variable, with target affect and voice shift type being treated as independent variables, which were tested separately. The model also included the interactions between target affect and shift type, with random intercepts for listener ID to account for individual differences. Holm-Bonferroni adjustments (Holm, 1979) were also applied to account for multiple comparisons. Confidence intervals were calculated using standard errors from the model output, and p values below .05 after the Holm-Bonferroni adjustment were considered statistically significant. Reference categories used for the analysis were neutral for the target affect and combined for the shift type. Model coefficients represent the log-odds of stimuli moving to a higher rating category on the ordinal scale (high activation affect categories), with negative values indicating a move towards the negative end of the scale (low activation categories).

Similar to experiment 1, an overall model could not be fitted across all scales at once due to the nature of the experiment (i.e., only two of the target affects were available as perceived affects for each task), which caused numerical issues and quasi-complete separation of data during modelling. The per-scale model was therefore considered the most suitable and robust for the analysis presented. Since the experiment was conducted with a single measure to capture both affect identification and intensity rate, the model was written so that all significant values shown in the tables and figures indicate significance on both of those dimensions. Further clarification is given below. Results are presented in figures 70-75, and tables 18-20.

Figure 70 presents a heatmap confusion matrix showing listener's judgements of affect per scale for each target affect and voice shift type, based on a binary identification of affect (e.g., for the *sad-happy* scale, any positive values are counted as one instance of *happy* as the perceived affect, and negative values count as an instance of *sad*, and so on).

Figure 71 presents an alternative representation of affect recognition rates for each affect-targeting triad of stimuli, also based on binary preprocessing. Figures 72-74 show the results of the ordinal regression analysis as confidence intervals over the coefficient estimates for the effects of target affect, voice shift type and their interactions for each test scale. Finally, figure 75 shows the confidence intervals for each triad of affect-targeting stimuli on each scale, plotting the confidence intervals levelled to the original intensity scale of the perception test instead of in terms of coefficient estimates.

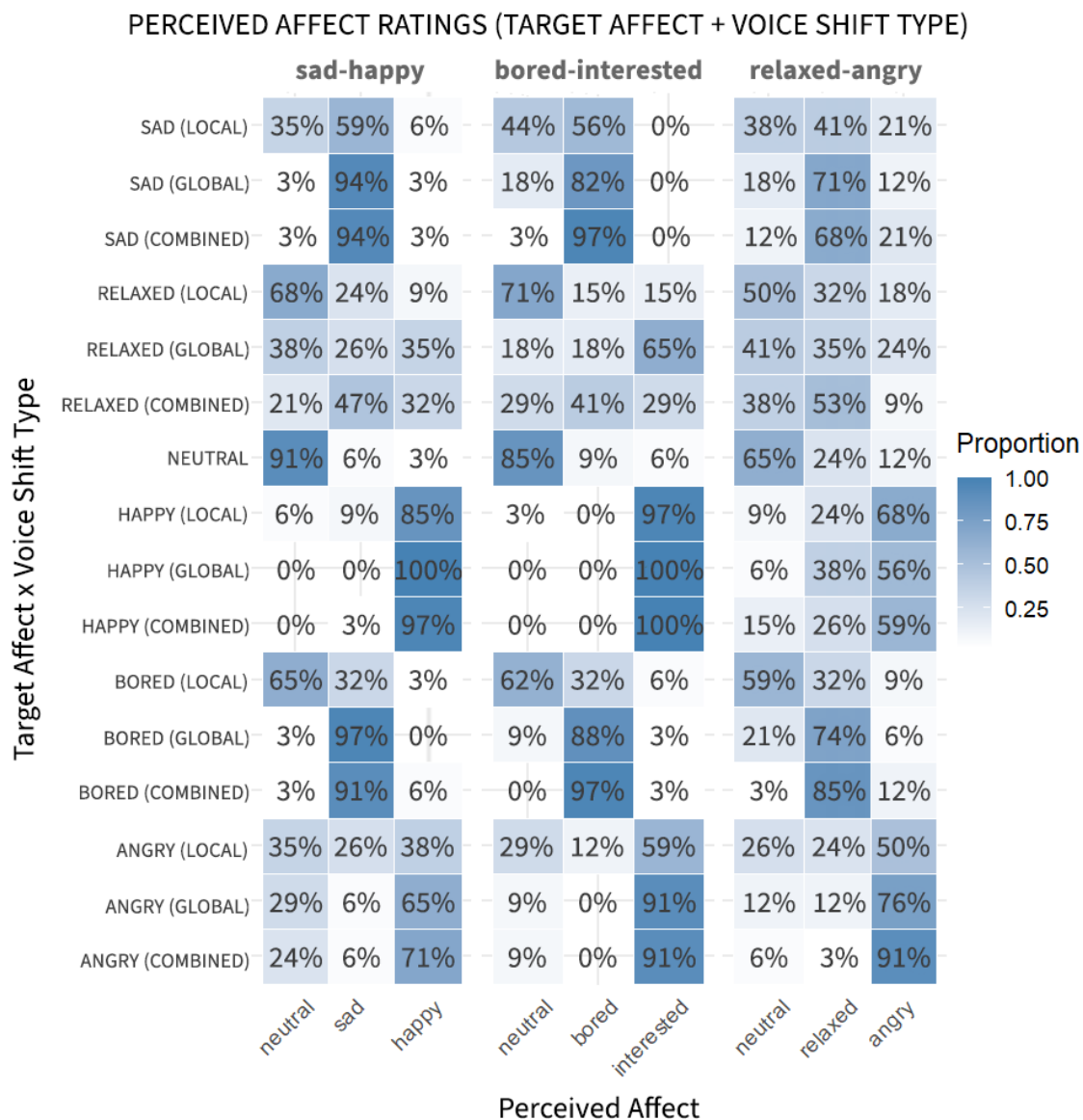


Figure 70 - Affect identification heatmap confusion matrix showing the percentages of affect identification rates (perceived affect) based on target affect and voice shift types, and on binary identification of affect (i.e., considering only categorical identification on each end of the test scales, without the effects of intensity scores).

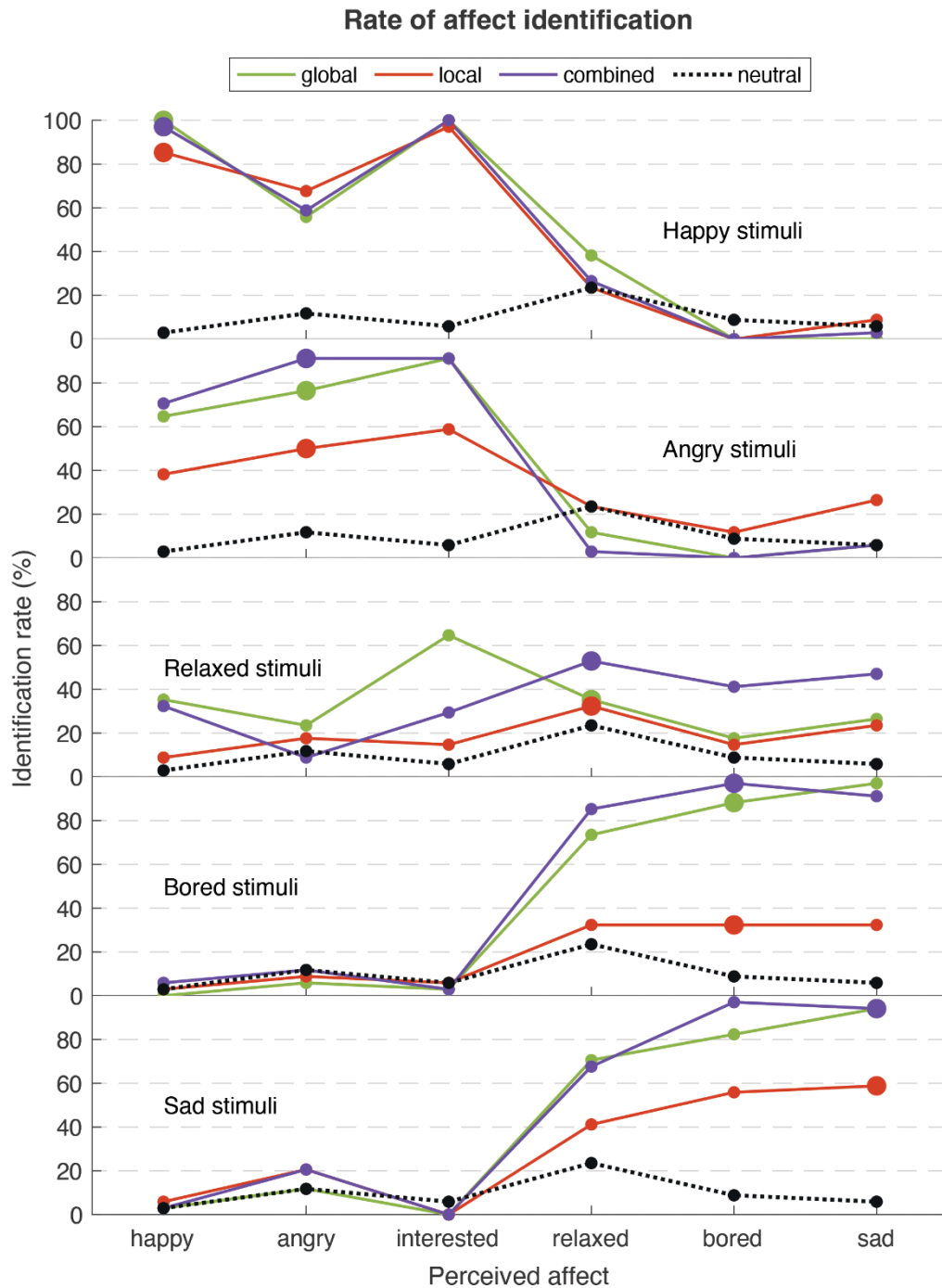


Figure 71 – Alternative representation of affect identification rates for each triad of affect-targeting stimuli. Titles by each section of the image indicate the stimuli’s target affect. Affect listings on the horizontal axis represent perceived affect from listeners, and the vertical axis represents percentages of identification rates. Identification rates are based on binary identification of affect (i.e., considering only categorical identification on each end of the test scales, without the effects of intensity scores). Neutral is shown as the dotted black line for comparison. The larger dots show the values of the target affect for each panel.

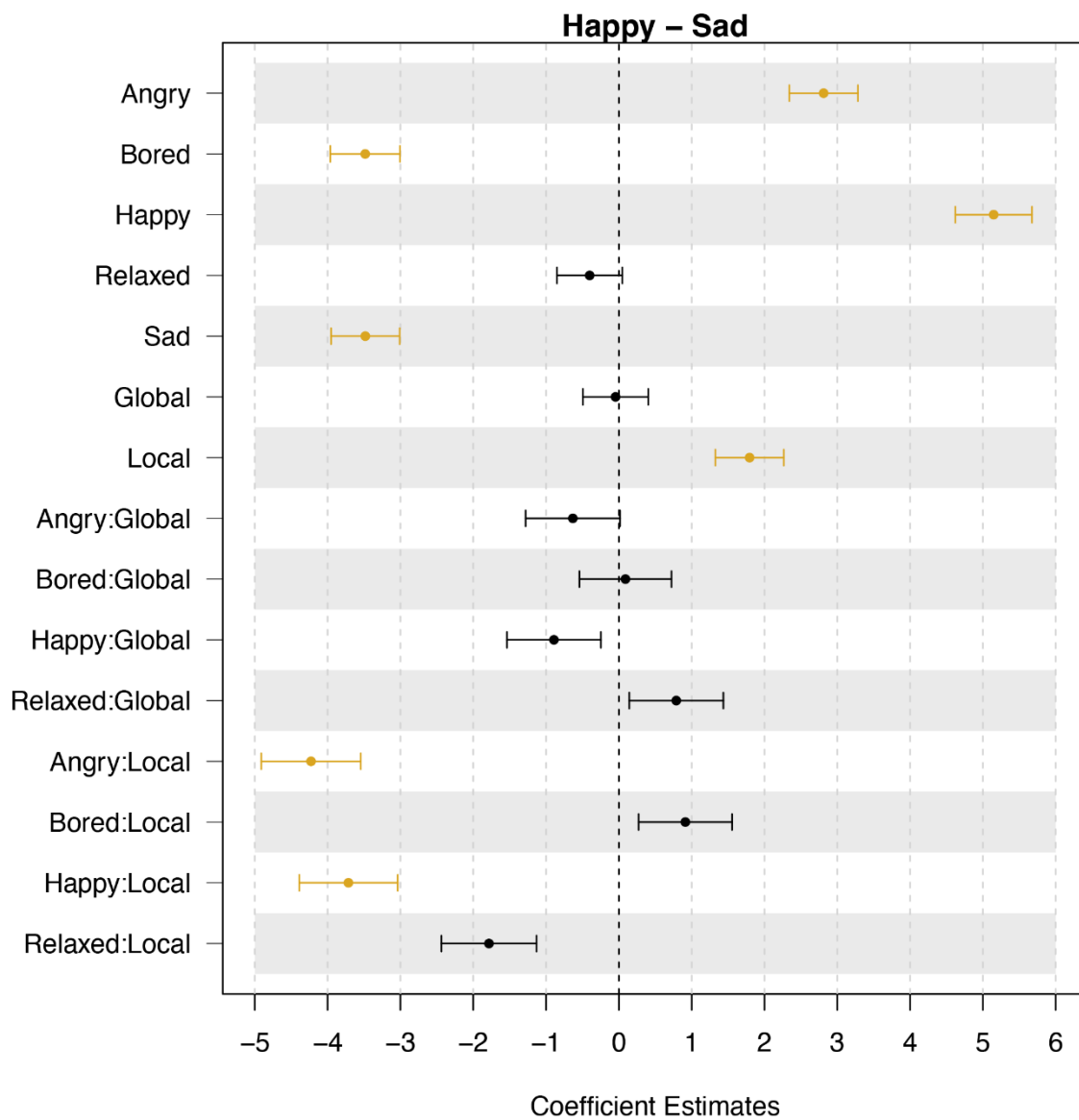


Figure 72 – Coefficient estimates for target affect, shift type and interaction effects from the mixed-effects ordinal regression analysis results for the *sad-happy* scale. Yellow lines indicate Holm-Bonferroni adjusted significances. Sidebars indicate Standard Error to each side. Bars completely above 0 on the horizontal axis indicates positive correlation, bars intersecting 0 indicate no correlation, and bars below 0 indicate negative correlation (see table 18 for numerical values).

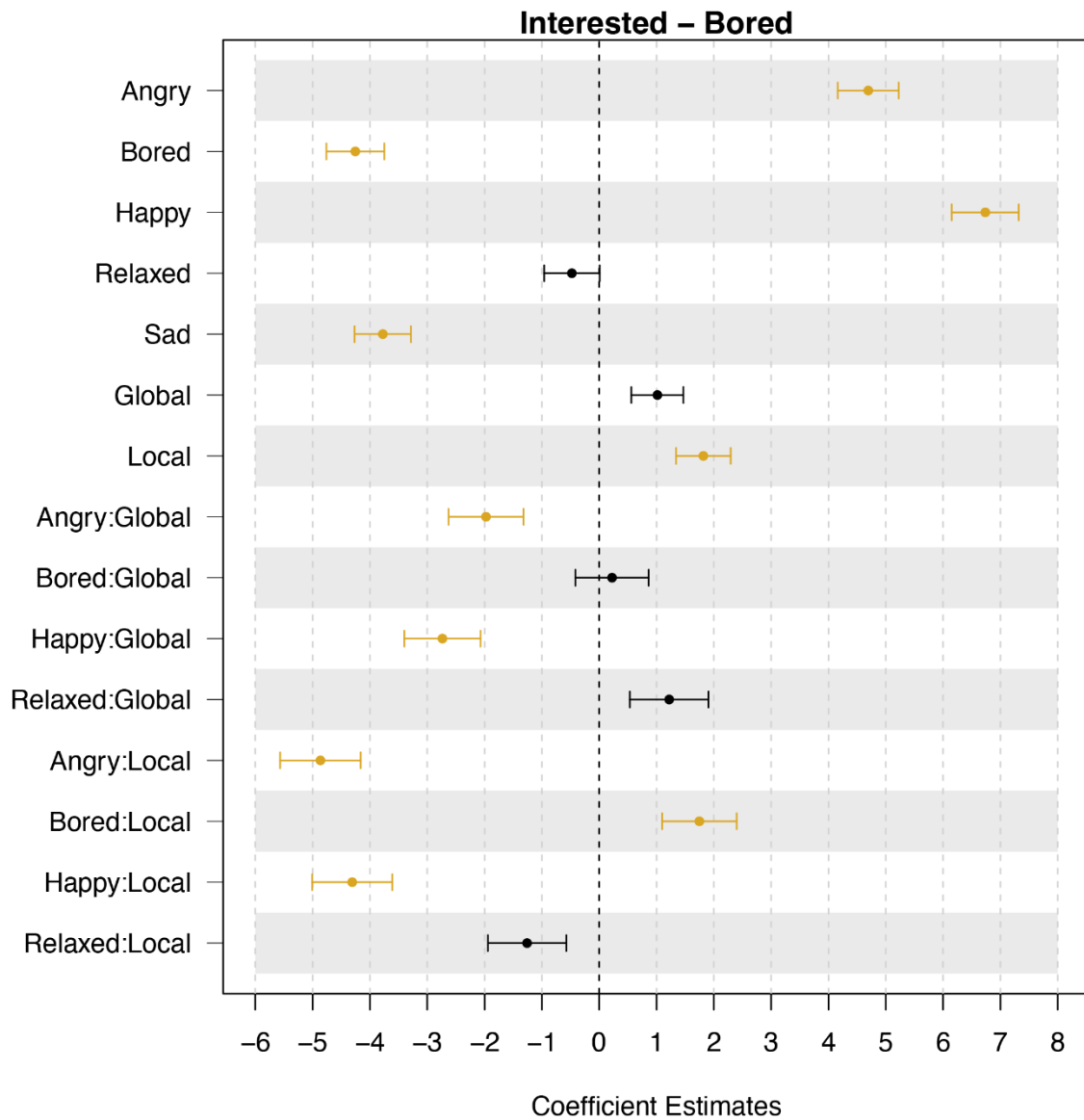


Figure 73 – Coefficient estimates for target affect, shift type and interaction affects from the mixed-effects ordinal regression analysis results for the *bored-interested* scale. Yellow lines indicate Holm-Bonferroni adjusted significances. Sidebars indicate Standard Error to each side. Bars completely above 0 on the horizontal axis indicates positive correlation, bars intersecting 0 indicate no correlation, and bars below 0 indicate negative correlation (see table 19 for numerical values).

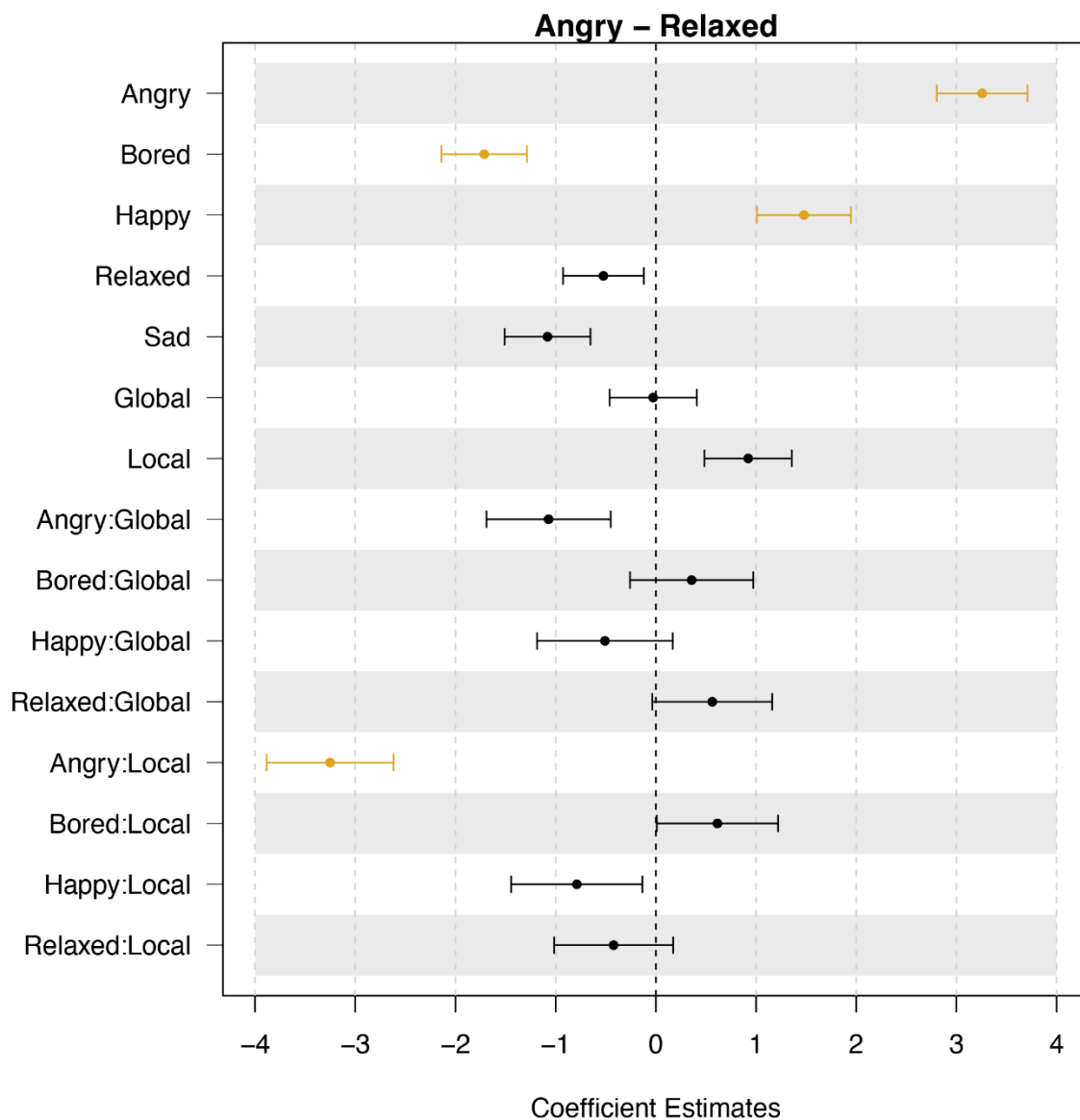


Figure 74 – Coefficient estimates for target affect, shift type and interaction effects from the mixed-effects ordinal regression analysis results for the *relaxed-angry* scale. Yellow lines indicate Holm-Bonferroni adjusted significances. Sidebars indicate Standard Error to each side. Bars completely above 0 on the horizontal axis indicates positive correlation, bars intersecting 0 indicate no correlation, and bars below 0 indicate negative correlation (see table 20 for numerical values).

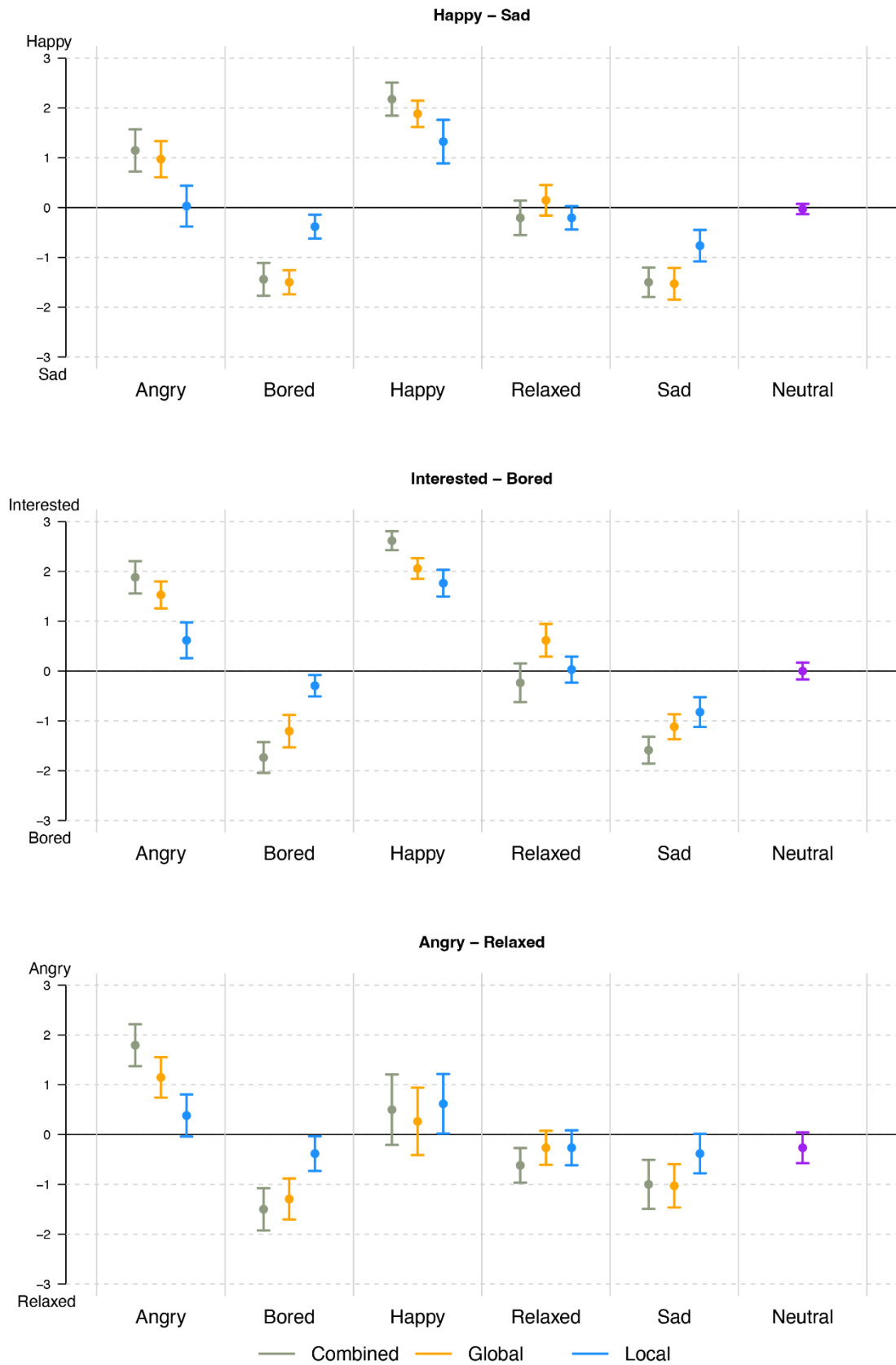


Figure 75 – Intensity score confidence Intervals for all target affects and triads of voice shifts across all scales. Error bars indicate 95% confidence intervals (± 2 standard errors) around each mean.

Table 18 – Ordinal logistic regression results for the *sad-happy* scale with Holm-Bonferroni adjustments, showing main effects of target affect, voice shift type and interactions on listeners’ perception of affect intensity. Positive estimates indicate positive correlation/above-chance performance relative to the reference categories; negative estimates indicate negative correlation/below-chance performance. Values close to 0 indicate no significant effect. Reference categories are *neutral* (target emotion) and *combined* (shift type). Absent interactions were dropped due to rank deficiency from showing patterns too similar to the reference levels. Significance levels: * $p < .05$, ** $p < .01$, *** $p < .001$.

Target Emotion × Voice Shift Type Effects on Intensity Ratings (Sad-Happy Scale)					
Predictor	Estimate (β)	SE	Z-value	P-value	P-value (Holm-Adjusted)
Main Effects					
Target Affect: Angry	2.81	0.47	5.96	<.001	<.001***
Target Affect: Happy	5.15	0.53	9.80	<.001	<.001***
Target Affect: Bored	-3.48	0.48	-7.28	<.001	<.001***
Target Affect: Sad	-3.48	0.47	-7.38	<.001	<.001***
Target Affect: Relaxed	-0.40	0.45	-0.90	.37	1.00
Voice Shift: Global	-0.05	0.45	-0.10	.92	1.00
Voice Shift: Local	1.80	0.47	3.82	<.001	.001**
Interactions					
Angry vs. Global	-0.63	0.65	-0.97	.33	1.00
Bored vs. Global	0.09	0.63	0.14	.89	1.00
Happy vs. Global	-0.89	0.65	-1.38	.17	1.00
Relaxed vs. Global	0.79	0.65	1.22	.22	1.00
Angry vs. Local	-4.23	0.68	-6.18	<.001	<.001***
Bored vs. Local	0.91	0.64	1.43	.15	1.00
Happy vs. Local	-3.71	0.68	-5.49	<.001	<.001***
Relaxed vs. Local	-1.78	0.65	-2.73	.01	.05*

Table 19 – Ordinal logistic regression results for the *bored-interested* scale with Holm-Bonferroni adjustments, showing main effects of target affect, voice shift type and interactions on listeners’ perception of affect intensity. Positive estimates indicate positive correlation/above-chance performance relative to the reference categories; negative estimates indicate negative correlation/below-chance performance. Values close to 0 indicate no significant effect. Reference categories are *neutral* (target emotion) and *combined* (shift type). Absent interactions were dropped due to rank deficiency from showing patterns too similar to the reference levels. Significance levels: * $p < .05$, ** $p < .01$, *** $p < .001$.

Target Emotion × Voice Shift Type Effects on Intensity Ratings (Bored-Interested Scale)

<i>Predictor</i>	<i>Estimate (β)</i>	<i>SE</i>	<i>Z-value</i>	<i>P-value</i>	<i>P-value (Holm-Adjusted)</i>
<i>Main Effects</i>					
Target Affect: Angry	-1.98	0.66	-3.02	.003	.02*
Target Affect: Happy	6.74	0.59	11.52	<.001	<.001***
Target Affect: Bored	-4.26	0.51	-8.43	<.001	<.001***
Target Affect: Sad	-3.78	0.49	-7.67	<.001	<.001***
Target Affect: Relaxed	-0.48	0.48	-0.99	.33	.65
Voice Shift: Global	1.02	0.45	2.24	.03	.13
Voice Shift: Local	1.82	0.48	3.81	<.001	.001**
<i>Interactions</i>					
Angry vs. Global	-1.98	0.66	-3.02	.003	.02*
Bored vs. Global	0.22	0.64	0.35	.73	.73
Happy vs. Global	-2.74	0.67	-4.11	<.001	<.001***
Relaxed vs. Global	1.22	0.69	1.78	.08	.26
Angry vs. Local	-4.86	0.70	-6.93	<.001	<.001***
Bored vs. Local	1.75	0.65	2.69	.01	.04*
Happy vs. Local	-4.31	0.70	-6.17	<.001	<.001***
Relaxed vs. Local	-1.26	0.68	-1.84	.07	.26

Table 20 - Ordinal logistic regression results for the *relaxed-angry* scale with Holm-Bonferroni adjustments, showing main effects of target affect, voice shift type and interactions on listeners' perception of affect intensity. Positive estimates indicate positive correlation/above-chance performance relative to the reference categories; negative estimates indicate negative correlation/below-chance performance. Values close to 0 indicate no significant effect. Reference categories are *neutral* (target emotion) and *combined* (shift type). Absent interactions were dropped due to rank deficiency from showing patterns too similar to the reference levels. Significance levels: * $p < .05$, ** $p < .01$, *** $p < .001$.

Target Emotion × Voice Shift Type Effects on Intensity Ratings (Relaxed-Angry Scale)					
Predictor	Estimate (β)	SE	Z-value	P-value	P-value (Holm-Adjusted)
Main Effects					
Target Affect: Angry	3.26	0.45	7.19	<.001	<.001***
Target Affect: Happy	1.48	0.47	3.15	.002	.02*
Target Affect: Bored	-1.71	0.43	-4.02	<.001	<.001***
Target Affect: Sad	-1.08	0.43	-2.53	.01	.16
Target Affect: Relaxed	-0.52	0.40	-1.30	.20	1.00
Voice Shift: Global	-0.03	0.43	-0.06	.95	1.00
Voice Shift: Local	0.92	0.44	2.11	.04	.35
Interactions					
Angry vs. Global	-1.07	0.62	-1.73	.08	.76
Bored vs. Global	0.36	0.62	0.58	.56	1.00
Happy vs. Global	-0.51	0.68	-0.75	.45	1.00
Relaxed vs. Global	0.56	0.60	0.94	.35	1.00
Angry vs. Local	-3.25	0.63	-5.14	<.001	<.001***
Bored vs. Local	0.61	0.61	1.02	.31	1.00
Happy vs. Local	-0.79	0.66	-1.21	.23	1.00
Relaxed vs. Local	-0.42	0.59	-0.71	.48	1.00

An interpretation of the results above is here presented. The ordinal regression analyses revealed significant main effects and interactions across all three scales (*sad-happy*, *bored-interested* and *relaxed-angry*), indicating that both variables (target affect and voice shift type) influenced listeners' perception of affect.

In terms of main effects of target affect categories (presented in the upper halves of tables 18-20 and figures 72-74), all target affects showed strong and consistent effects

relative to the neutral stimuli across all scales. The *happy*-targeting stimuli consistently produced the largest positive coefficients towards the high activation affects on each scale (*sad-happy*: $\beta = 5.15$, *bored-interested*: $\beta = 6.74$, *relaxed-angry*: $\beta = 1.48$), indicating that listeners consistently rated these stimuli on the higher end of the high activation sides of the scales. Conversely, low activation affects (*bored* and *sad*) consistently showed large negative coefficients, indicating successful affect cueing in terms of activation categories.

For overall voice shift types, the local shifts showed significant positive effects on the *sad-happy* ($\beta = 1.80$, $p < .001$) and on the *bored-interested* ($\beta = 1.82$, $p < .001$) scales, suggesting that, considering the effects of the local shifts across all affects, local shifts were more effective than the combined shifts for these two scales. Here, combined performed better than global ($\beta = -0.03$), considering that the estimate values are taken using the combined category as the reference level. In this analysis, global shifts showed minimal main effects across all scales relative to combined (*sad-happy*: $\beta = -0.05$, *bored-interested*: $\beta = 1.02$, *relaxed-angry*: $\beta = -0.03$).

Based solely on the binary identification rates in figures 70-71, all the affects tested seem to have achieved very high rates of identification for at least one of the stimuli in each triad. The most effective in achieving correct target affect identification were the combined shifts, with identification rates ranging from 53% for *relaxed* to 94% for *sad*. The global stimuli were also quite effective: for *happy*, the global stimulus achieved 100% identification on its respective scale, with the *sad* local stimulus matching the *sad* combined (94%). Overall, the local seemed to perform much worse than the combined and global stimuli, which goes directly against the results seen for the main effects of the voice shift types; therefore, to account for this difference, an analysis of the interaction effects is presented in further detail.

Interaction effects estimates and p values can be seen in the bottom halves of tables 18-20 and in the accompanying figures 72-74. Additionally, figure 75 shows the confidence intervals of scores for each target affect and voice shift type per-scale. From these, we can immediately see a direct contrast with the main effects results, especially for the high activation affects: interaction effects show that, for *angry* and *happy*-targeting stimuli, the local shifts were significantly less effective than the combined.

The local *angry* stimulus shows large negative coefficients across all scales (*sad-happy*: $\beta = -4.23$, *bored-interested*: $\beta = -4.86$, *relaxed-angry*: $\beta = -3.25$, all $p < .001$), indicating poorer performance in identification and intensity ratings than the combined baseline and the global shifts. This pattern is seen in figures 72-74, which clearly show the local *angry* as performing significantly below the global *angry* stimuli, and in figure 75 in terms of the intensity confidence intervals for the combined and global *angry* stimuli. This pattern also mirrors the identification percentages seen in figure 70, further confirming that this was both an effect of target identification and perceived intensity. On the *sad-happy* scale, this may suggest that the local *angry* was more well-differentiated from the *happy* perceived affect than the other two voice shifts, as indicated by the confidence intervals in figure 75.

For the global *angry* stimulus, the only significant effect is found on the *bored-interested* scale ($\beta = -1.98$, $p = .018$), indicating significant lower intensity in its perception and a shift towards the perception of boredom compared to the combined *angry*. While this appears inconsistent with the binary identification rates in figures 70-71 and even the distance between the combined and global *angry* confidence intervals in figure 75, this significant effect is likely to come from an important feature in the statistical analysis: the smaller standard error bars for these two stimuli in the *bored-interested* panel of figure 75 indicates a more meaningful separation of ratings; i.e., here, the deviation of values for this stimulus is less dramatic than in the other scales, therefore resulting in the small significance found. In summary, for the *angry* affect, the combined stimuli were the most effective both in terms of accurate target affect identification and activation category, followed by the global (which was only significantly different from the combined on the *bored-interested* scale), and with the local consistently drifting closer towards the low activation ends of the scales.

As for the *happy*-targeting stimuli, results indicate the following. The local *happy* stimulus, like the local *angry*, also shows negative interactions on the *sad-happy* ($\beta = -3.71$, $p < .001$) and *bored-interested* ($\beta = -4.31$, $p < .001$) scales, which also indicates a shift towards the low activation end of these two scales in relation to the combined *happy* stimulus. We can see in figure 75 that these negative effects come from lower intensity ratings assigned to the local *happy* stimulus, considering how these results contrast with the extremely high binary identification rates of *happy* in figure 70. On the *relaxed-angry*

scale, the local *happy* did not show significant difference from the combined. Other significant values for the *happy* stimuli were only found for the *happy* global stimulus on the *bored-interested* scale ($\beta = -2.74, p < .001$), following a similar pattern to the *angry* global and indicating that, for the *happy* affect, no significant differences were found the intensity scores and activation category identification between the global and combined stimuli for the *sad-happy* and the *relaxed-angry* scales.

The last significant interaction found was for the *bored* local stimulus on the *bored-interested* scale ($\beta = 1.75, p = .04$), the positive value indicating a significant drift of this stimulus away from the *bored* end of the scale. As seen in figure 75, this was likely a result of lower intensity scores assigned to this stimulus. No other *bored* stimuli achieved any significance across the other scales, indicating little differentiation from the baseline neutral and combined shifts in terms of identification and intensity. As mentioned, the interactions for the *sad*-targeting stimuli were automatically dropped by the statistical model due to rank deficiency, which occurs when a statistical model has too many parameters to reliably estimate given the patterns in a dataset. This practically means that interactions for the *sad* stimuli were too similar to other interactions already in the model (likely the *bored* stimuli), making it unable to distinguish between these patterns and dropping coefficients to prevent overfitting. What this suggests is that *sad* and *bored* stimuli had nearly identical response patterns and, therefore, because most of the *bored* interactions presented no statistical significance, there were also no statistically significant interactions for *sad*. This is confirmed by the *bored-interested* scale in figure 75 – the confidence interval for the local *bored* is much more differentiated from the global and combined *bored* than the pattern for the *sad*-targeting shifts shows, indicating that even this one significant finding for *bored* might not have been reflected in the *sad* stimuli.

The *relaxed-angry* scale showed the fewest significant effects overall, with only the *angry* local stimulus reaching significance after the Holm-Bonferroni correction. Like the previous experiment, the *relaxed* stimuli remained the closest to and least differentiated from the neutral baseline, showing only a maximum of 53% identification for the combined *relaxed* as its target affect on its respective scale (in binary classification terms) and a slightly stronger effect in its rating as *interested* on the *bored-interested* scale (figure 75).

While these results support the hypothesis that the combined stimuli would be the most effective, it was unexpected to see that, in all cases, the global performed better than the local stimuli. The global *relaxed* was often perceived as the *interested* affect, possibly due to its positive valence.

The apparent contradiction between the significant positive main effects of local shifts and the negative interaction coefficients for specific target affects highlights the critical importance of examining interaction effects alongside main effects. While local shifts demonstrated significant positive main effects across multiple scales (e.g., $\beta = 1.80$, $p < .001$ on the *sad-happy* scale), indicating that local manipulations are generally more effective than combined approaches when averaged across all target affects, the significant negative interactions reveal important exceptions to this general pattern. Specifically, the large negative coefficients for the local *angry* and the local *happy* (ranging from $\beta = -3.25$ to $\beta = -4.86$) indicate that local shifts are substantially less effective in cueing high activation affects compared to the combined shifts. This pattern suggests that, statistically speaking, local shifts work relatively well for cueing most affects (*bored*, *sad* and *relaxed* especially) creating an overall positive main effect, but fail dramatically for high activation affects, creating strong negative deviations from this general trend. The interaction effects are arguably more theoretically and practically significant than the main effects, as they identify specific conditions where the generally effective local manipulation strategy becomes ineffective, preventing overgeneralization of the positive main effect and suggesting a need for affect-specific voice source manipulation approaches.

In terms of affect differentiation and unique cueing, results can be interpreted as follows. Most significant values were found on the *bored-interested* scale, which also shows an interesting pattern when looking at figures 70-71: the high activation affects *happy* and *angry* were often perceived as *interested*, sometimes to an even higher degree/intensity than their perception as their target affect categories on their respective scales. The results for the *happy* local stimulus on the *sad-happy* ($\beta = -3.71$, $p < .001$) and *bored-interested* ($\beta = -4.31$, $p < .001$) scales, in contrast to its lack of significant effect on the *relaxed-angry* scale, indicate that the *happy* affect was well-differentiated from *angry*, especially considering these values against the combined baseline (meaning that the combined stimuli achieved even higher unique cueing). However, following the same logic and looking at

figures 70-71, neither the *angry* nor the *happy* stimuli appear to be well differentiated from their cueing of *interested*.

The overlap for both the *happy* and *angry* stimuli recognized as *interested* might arise from the speculation that, one might argue, both happiness and anger intrinsically imply a high level of interest, whether in the full message or in specific parts of the utterance. As the *interested* recordings were discarded due to focus shifts, it is perhaps not too farfetched to assume that the high deviations from the neutral baseline, especially for the local and combined stimuli, might have conjured the *interested* affect for listeners.

Regarding differentiation within the low activation categories, the responses to the stimuli targeting the *bored* and *sad* affects are virtually identical, indicating that the two affects were essentially perceived as one or the other, without much specification between them. Additionally, those stimuli were also often perceived as *relaxed*, with no significant differences between voice shift types in any of the scales presented. As mentioned, the fact that the interactions of the *sad* stimuli were dropped by the statistical model due to rank deficiency caused by the presence of patterns too similar to other interactions in the dataset further confirms this, and we speculate that these very similar interactions might have indeed been those that included the *bored* stimuli. This indicates that the cueing used in these stimuli were not affect-specific. In fact, the *bored* and *sad* stimuli were more successful in cueing the *relaxed* affect than the *relaxed*-targeting stimuli.

It may also be important to note that non-significant interactions have a meaning of its own in the context of this experiment. They may indicate that, for the low activation effects, the proposed affect-related voice shifts are only consistently cueing their activation category, and that different parameters, more dramatic modifications or additional information channels may be required to achieve differentiation between target affect categories. For example, the effects of these dramatic parameter shifts are seen in the results for the *happy* stimuli, which have the largest deviation from the neutral baseline.

In summary, the results suggest that the stimuli targeting low activation states remain ambivalent as they were equally perceived as *sad*, *bored*, and, in some cases, *relaxed*, even with the shift in the neutral baseline. The *relaxed* stimuli remain elusive –

there seems to be some differentiation in terms of valence and the combined *relaxed* seems to have been effectively recognized, but the analysis still reflects its closeness to the neutral.

The high activation states achieved much stronger interaction effects. *Happy* remains the most effective in terms of statistical significance, being well differentiated from *angry*. While stimuli targeting *interested* were not included in this batch for reasons mentioned previously, the overlap between the *interested* affect and the *angry* and *happy* stimuli may indicate that the high local contour shifts of the two high activation affects were sufficient to evoke the interested affect, to a degree where no significant difference was found between its cueing and that of some stimuli's targets. A broader perception of *interested* might potentially be induced without the imbued overtones of *happy* or *angry*, and possibly without the focus shifts seen in the production data. Investigating this would, however, require an independent modelling of *interested* either based on the literature, on the presented results, or on a new set of production data.

6.4. Discussion and Conclusions

This experiment has confirmed our initial hypothesis in the sense that the strongest signaling of each affect was achieved by the combined stimulus that targeted that affect. Results also seem to indicate that, particularly for the *happy*-targeting stimuli, the effects of global and local manipulations are additive, given the negative coefficients for the *happy* stimuli relative to the combined and confidence interval seen in figure 75.

However, what came out as a surprise were the relatively low strength ratings for the local stimuli, which contrasts with findings in earlier studies. This might be a result of the change of baseline from the recorded neutral, which is now closer to the high activation affects than the previous experiment. While recognition rates are higher than in the previous experiment, the change in methodology (discrimination vs. scale-based tasks) makes it difficult to make direct comparisons, as listeners only judged the stimuli between two perceived affect categories at a time. Additionally, the intensity scales raise another question – in contrast to the other emotions cued in this study, how can one rate the intensity of *relaxed* as an affect that, in essence, means the absence of intensity?

Finally, these results add to earlier findings that suggest that affect cueing is often ambivalent, i.e., some specific voice source characteristics are associated with multiple

affects rather than there being a one-to-one mapping between voice quality and affect. The next set of tests will therefore add visual context cues to the vocal stimuli to investigate how the added context affects the differentiation between low activation affects (or between high activation affects), or how much a visual context that contrasts with the vocal stimulus interacts and potentially distorts a listener's perception of vocal stimuli targeting specific affects.

7. PERCEPTION EXPERIMENTS 3 AND 4 – THE VISUAL CHANNEL

This second phase of the research aims to investigate the effects of the visual context on the resynthesized vocal stimuli. With the results from the vocal-only experiments and a set methodology for creating the vocal stimuli, the visual experiment was designed. Additional required resources and background are detailed below.

First, a method of creating facial expressions that could be easily modified and refined was desired, similar to how the vocal stimuli were created over multiple iterations and adjustments. Because of that, it was decided, rather than to hire and photograph actors or manually illustrate portraits, to utilize the free 3D software Daz Studio (Daz Productions, 2022), chosen due to its approachable interface (in comparison to other 3D modelling software), in-software animation capabilities, and the immediate availability of high-quality, pre-rigged realistic human 3D models. These models allow for manipulation of facial movements through a wide range of sliders, enabling modifications of features in a continuous and measurable manner which made it possible to create expressions ranging from very intense to very subtle.

The second consideration involves a scientific methodology for creating the facial expressions, rather than utilizing only personal experience and visual instinct of what certain expressions may look like. For this, Paul Ekman Group's Facial Action Coding System (FACS) (Ekman et al., 2002), which investigates and quantifies shifts in facial muscle movement that create facial expressions, was used. The FACS does so by assigning scores to numbered Action Units (AUs) of the face, which "represent the muscular activity that produces momentary changes in facial appearance" (Ekman et al., 2002, p. 13). It also quantifies the intensity of the gesture: neutral receives a score of 0 in the system, which then starts from *trace* (A score), defined as the minimum shift necessary for the viewer to identify that a change from a neutral expression has happened, to *maximum* (E score), which is determined by the anatomical limit of the muscle. For example, a lowering of the brow is classified as AU4, and therefore a moderate lowering of the brow would be defined as a score of 4C. Because the exact appearance of facial expressions changes depending on the facial structure (e.g., bone structure, musculature, fatty deposits, shape features and condition of skin) of an individual's face, the FACS focuses on common changes that occur

in these specific Action Units, i.e., the movement of the muscle under the skin which results in the facial expressions we see. Each target affect is here described in terms of these AUs.

Because the FACS mainly describes the muscular actions themselves without directly connecting them to facial expressions of emotion, an additional practical reference was also used for the application of the FACS to the 3D model. The work of Zarins (2017) brings together the Action Units and scores of the FACS into discrete emotion sets, providing visual, muscular and surface level references targeted towards sculptors and artists with various alternatives of expressing a certain affect (e.g., happiness expressed either with a closed mouth smile, a full smile, or laughter). This allowed for a better transference of the terminology of the FACS into the 3D software used for creating the stimuli, given the fact that the 3D rigging only allows for manipulations at a surface level. It also bears mentioning that there is no single formula for each emotion to be expressed in the face – people may use different intensities of an AU, or even a different combination of AUs. For the purpose of this research, the most stereotypical descriptions in Zarins (2017) used in realistic sculpting were followed, as they consist of long-established conventions used to visually convey target emotions to viewers. Further detail on the creation of the visual stimuli is given in the materials section below.

7.1. Materials

This section details the materials used in the present experiment, including the vocal stimuli and the methodology for the creation of the visual stimuli.

7.1.1. Vocal Stimuli

The vocal stimuli used for this test were the same as those used in the ABAIR neutral baseline perception test in terms of global and local shifts, detailed in chapter 6. Tables 10-11 show the global and local shifts, respectively. Tables 15-17 show the Mean Absolute Difference values extracted to test the VSG system's performance and the parameter differences of each stimulus to the neutral baseline. Figures 67-69 show examples of the

VSG system's output contour for the *angry* combined stimulus; finally, other affects, transformation types and parameters can be found in Appendix D¹¹.

7.1.2. Visual Stimuli

The creation process of the visual stimuli is detailed below, followed by a summary of the final set of manipulations to the facial action units for each target affect used in the perception test.

The creation of these facial stimuli is guided primarily by facial expression documentations done by English-speaking researchers (the FACS), as there are no studies that specifically investigate how Irish speakers perceive facial expressions. Based on evidence for the universality of facial expressions across cultures (Ekman, 1989), it is assumed that visual channel characteristics overlap across English and Irish speakers (and potentially most Western European cultures), given the small geographical distance between these cultures. To start, a set of facial movements deviating from the neutral for each affect based on the FACS and Zarins (2017) had to be determined, and is described below for each emotion presented in this project.

Ekman et al. (2002) defines the neutral face as when there is 'no action' of any of the muscles, which includes wrinkles, pulling, muscle budes, drooping or any changes caused by alterations in the subject's emotional state. This also includes emotions that would cause the body to be in a state of lower activation such as boredom or tiredness, which cause muscles to relax beyond the normal levels of the average neutral state. The same is exemplified in Zarins (2017), where the neutral expression is described as the state in which the muscles of the face are relaxed and there is no active pulling or tension being applied.

Figure 76 provides an example of the neutral face from the FACS Manual, and an illustration of the terminology used to describe the facial movements used for creating these stimuli.

¹¹ Parameter illustrations in figures 62-64 and in Appendix D were created through the custom Praat script by Rodgers (2022) and edited through Adobe Photoshop (Adobe Inc., 2023) for formatting adjustments.

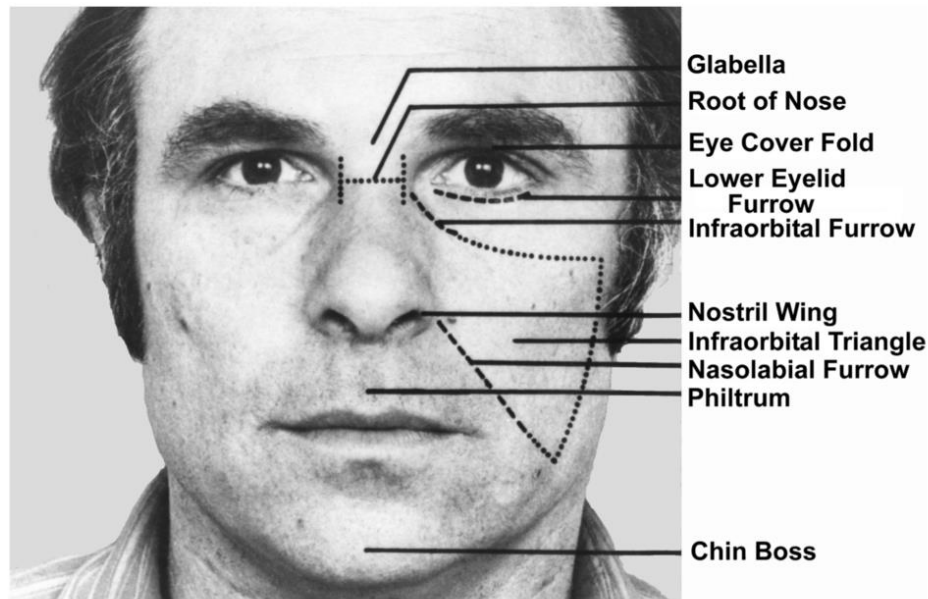


Figure 76 - Illustration of the neutral face with names and locations of facial areas.
Extracted from Ekman et al. (2002).

Angry

The following action units were chosen based on a stereotypical expression of anger detailed in Zarins (2017, p. 195). As mentioned, in many reference images of actors in the same body of work and other pieces of literature (e.g., Keltner, Sauter, et al., 2019), the action units and intensity scores stated to represent expressions of anger may vary. For example, anger, in contrast to rage, doesn't always involve mouth opening action. Therefore, to keep the expression subtle for experimental purposes, it was decided to keep the mouth closed with pressed lips.

The first action unit chosen to create an angry facial expression is AU4, named *brow lowerer*. This consists of a lowering of the inner and central portions of the eyebrow. This also pulls the eyebrows closer together, resulting in wrinkling between the eyebrows, depending on intensity.

AU5, called *upper lid raiser*, pulls the upper lid back into the eye socket, making the eye aperture wider. While AU5 was used in this set of expressions, anger on the face can also go the opposite direction with AU7 (*lid tightener*) which tightens the eyelids instead of opening them, making the eyes narrower. The combination of AU4+5 imposes limits to the

intensity of each of these action units. AU4 naturally pushes the eye cover fold downwards, which may narrow the eye aperture, while AU5 works to widen this aperture. The resulting aperture is a compromise between these two actions.

AU23, or *lip tightener*, makes the red part of lips looks narrower, creating the pressed lips expression. It may produce slight wrinkling of the chin boss. Finally, AU38, *nostril dilator*, dilates or flares the nostrils. It can be an indication of a stronger intake of air, which then creates increased subglottal pressure and powers the lungs/voice source for vocal expressions of anger.

Happy

Like anger, expressions of happiness can also appear in a variety of ways such as a closed or open mouth smile. Therefore, the same logic was imposed, and action units were chosen to intentionally create a subtle *happy* expression.

The first action unit chosen for this was AU6, called *cheek raiser and lid compressor*. It draws the skin from the temples and cheek towards the eyes by contracting the outer band of muscle around the eyes. As a result, the infraorbital triangle is raised and the cheek is lifted upwards, which then pushes the skin around the eye towards the eye socket. This can narrow the eye aperture, bag, or wrinkle the skin below the eye, along with other further changes depending on the intensity.

AU12, or *lip corner puller*, pulls the corners of the lips back and upwards, creating a smile. This also affects many other adjacent areas such as deepening the nasolabial furrow, raising of the infraorbital triangle, and potentially deepening the infraorbital furrow to a small degree. For relaxed expressions, the intensity score is kept low (12B), therefore not affecting adjacent areas as much (or as many) as a wider smile.

The combination of AU6 and AU12 is often enough for a more discrete but clear expression of happiness or joy, even more so when AU12 is at its full intensity (i.e., anatomical limit) without breaking the lips into an open mouth smile. In one example in Ekman et al. (2002), AU7 (lid tightener) is also identified, but, for slighter expressions, one may not require its active use, being rather a side effect of AU6. For the aims of this

experiment, the AUs responsible for a lip-parted smile were not used, as that would be too strong an expression, and lip movement was therefore limited to AU12.

Sad

Expressions of sadness can also vary between subtle changes to one's facial musculature to strong portrayals such as crying. Because the sad vocal stimuli were already a rather low activation version that did not include more extreme indications of sadness such as breaking voice from crying, subtlety was applied through the action units described below.

First, AU1 (*inner brow raiser*) uses the frontalis muscle to raise the inner corner of the eyebrow, pulling the inner portion on eyebrows upwards and giving them an oblique shape. This may also cause the skin in the center of the forehead to wrinkle slightly, depending on the intensity.

This emotion also uses AU4 (*brow lowerer*) to pull the eyebrows closer together, in a similar manner to the angry expression. The difference is that, when combined with AU1, AU4 only goes as far in intensity as 4B – the combination of the lowering of the brow with the raising of its inner corners makes the movement more directed towards the middle of the brow and upwards, rather than just the lowering and downward bunching of the brow seen in AU4 in isolation.

Finally, AU15, or *lip corner depressor*, pulls the corners of the lips down, somewhat stretching the lower lip horizontally and producing a frowning expression. When stronger, this action can also create some pouching of skin below the lip corners.

Bored

Not presented as a predetermined set in Zarins (2017) nor part of Ekman's basic emotions (Ekman, 1999) are the *bored* and *relaxed* affects. Therefore, the action units used to produce these expressions were taken from the work of Keltner, Sauter, et al. (2019), who also describe a wider array of facial expressions for emotion research based on the FACS.

Given its nature as a low-effort and low activation state, the facial expression of boredom does not deviate much from the neutral. The only facial movement described for

this emotion is AU43, or *eye closure*. In this context, it means a slight drooping of the eyelids rather than full eye closure, indicating an overall state of inaction of the body. While this AU is shared with the *relaxed* affect, *bored* appears to feature slightly increased activation of this AU (a score of 43B). This means that the eye closure is slight, but the eye opening is markedly less wide than usual as a main feature of this affect.

Another potential option explored in the expression of boredom created for this study was AU55/56 (*head tilt*), described as independent from the tilting of other areas of the body (torso, shoulders). Here, the head is slightly cocked to the left (AU55) or the right (AU56) in a relaxed manner, without requiring application of effort to the neck muscles (scoring 55/56B). Although the inclusion of AU55/56 was explored, this AU was considered more as a gesture than part of a facial expression, and excluded from the final set of stimuli as it contrasted too much with the other visual stimuli.

Relaxed

For the relaxed affect, the expression of contentment in Keltner, Sauter, et al. (2019), produced with AU12 and AU43, was used as reference.

AU12 (*lip corner puller*), as in the expression of happiness, pulls the corners of the lips back and upwards to create a smile. For relaxed expressions, the intensity score is kept low (12B at most) so that this unit does not affect the adjacent areas as much as a wider smile would.

Then, there is AU43 (eye closure), which is shared with the *bored* affect. For a *relaxed* expression, there is an even slighter drooping of the eyelids than seen for boredom, indicating an overall state of relaxation of the body. When combined with AU12, this creates a feeling of contentment, i.e., low activation happiness. Only a trace score (43A) is necessary, as to not make the eye opening markedly less wide than usual and only indicating a lower activation state than neutral.

Interested

In line with the inclusion of *interested* on the previous perception test, it was decided to include this facial expression in the set as a contrasting stimulus to *bored*. While also not

present in Zarins (2017), Keltner et al. (2016) defines the *interested* expression in terms of the following action units.

AU1 (*inner brow raiser*) is again used to raise the inner corners of the eyebrows, pulling its inner portions upwards. This is combined with AU2, or *outer brow raiser*, which pushes the lateral portion of the eyebrows outwards. This combination raises the eyebrows fully, giving an arched shape to the brow. It also causes the lateral portion of the eye cover fold to be stretched upwards, slightly increasing the opening of the eyes. The intensity of these movements depends on the level of interest shown by the subject.

AU12 (lip corner puller) is once again used to create a slight smile by pulling the corners of the lips upwards. For *interested*, much like the *relaxed* expression, the intensity score is kept low (12B), possibly even lower than relaxed (12A) given that interest not always is accompanied by smiling.

Translating the Action Units to the 3D Model

With the action units for each affect in hand, the visual stimuli were created in the free version of Daz3D (Daz Productions, 2022), using one of their 3D male models (Ty) available in the Genesis 9 package.

First, a baseline was needed to apply the shifts systematically. It was decided to utilize the default state of the 3D model to portray the neutral affect. This state matched Ekman's description of the neutral expression in the sense that the muscles that would be activated during facial expressions were kept relaxed (i.e., built-in tension sliders were set at 0%), while movements that could go two different ways (e.g., lip corner movements up for happiness vs. down for sadness) were kept at 50% (or 0%, whenever the slider scale allowed for negative values). As before, an informal viewing session was conducted to confirm whether this state of the 3D model was perceived as a neutral expression, which was confirmed.

Then, corresponding parameters to the action units listed above were determined. The model used has a wide range of movement and individual modifiable parameters that allowed for very fine control of muscle movements; however, not all parameters directly correspond to the AUs, and therefore some adaptation was required following the

reference photographs in Ekman et al. (2002), Zarins (2017) and Keltner, Sauter, et al. (2019) when creating this final set of expressions (e.g., adding lip corner movement to the angry and interested affects). Table 21 below shows the action units and their corresponding parameter sliders in Daz3D. Table 22 presents a summary of the Action Units per emotion in a feature matrix.

Table 21 - Action Units for each affect and corresponding parameter sliders in Daz3D (Genesis Model).

Action Unit	Daz3D Parameters
AU1 – Inner Brow Raiser (Sad, Interested)	Brow Inner Up (or Brow Up)
AU2 – Outer Brow Raiser (Interested)	Brow Up
AU4 – Brow Lowerer (Angry)	Brow Down + Brow Squeeze
AU5 – Upper Lid Raiser (Angry)	Eyelid Open-Close Upper
AU6 – Cheek Raiser & Lid Compressor (Happy)	Cheek Squint + Eye Open-Close Lower
AU12 – Lip Corner Puller (Happy, Relaxed, Interested)	Mouth Smile + Mouth Dimple
AU15 – Lip Corner Depressor (Sad)	Mouth Frown
AU23 – Lip Tightener (Angry)	Mouth Press
AU38 – Nostril Dilator (Angry)	Nasal Flare
AU43 – Eye Closure (Bored, Relaxed)	Eyelid Open-Close Upper

Table 22 - Feature matrix showing the Action Units for each affect.

Action Unit/Affect	Angry	Happy	Sad	Bored	Relaxed	Interested
AU1 – Inner Brow Raiser	No	No	Yes	No	No	Yes
AU2 – Outer Brow Raiser	No	No	No	No	No	Yes
AU4 – Brow Lowerer	Yes	No	No	No	No	No
AU5 – Upper Lid Raiser	Yes	No	No	No	No	No
AU6 – Cheek Raiser & Lid Compressor	No	Yes	No	No	No	No
AU12 – Lip Corner Puller	No	Yes	No	No	Yes	Yes
AU15 – Lip Corner Depressor	No	No	Yes	No	No	No
AU23 – Lip Tightener	Yes	No	No	No	No	No
AU38 – Nostril Dilator	Yes	No	No	No	No	No
AU43 – Eye Closure	No	No	No	Yes	Yes	No

Another point made in the literature that was considered when creating these stimuli was the evidence that static facial features may subject emotional faces to biases (e.g., Kraus, 2017; Olivola, Funk, & Todorov, 2014), as facial features are not static in day-to-day interactions and could potentially distract subjects once the perception test was conducted. Therefore, it was decided to create simple animations with lip-synching to match the vocal stimuli, along with some degree of eye movement to avoid having the character stare unnaturally at the listeners.

This was done through previous animation training and real-life references for the lip and eye movements, recorded while speaking the utterance '*Gheobhmid bá d ar an lá sin*'. Because high quality animation is a very time-consuming process, the animations were kept simple and systematic given the time constraints of the project but still made as natural as possible. Lip positions were created as key frames aligned with the middle point of each phoneme from the neutral ABAIR utterance. This alignment was based on a base framerate of 30 frames-per-second, calculated by multiplying the timing (in seconds) of the phoneme by 30 to find the position of each key frame in the animation timeline. In-between frames were filled in through the animation interpolation algorithm in Daz3D, and adjusted in case any transitions were unnatural or required additional key frames. The same was done with the eye movement, which was kept much simpler.

For the purposes of replicability, the animation parameters are listed in table 23, organized according to the muscle movements on the sliders relative to each key frame. These were kept uniform across all affects, only changing when they overlapped with facial parameters used specifically for the facial expressions, through addition or subtraction of percentages.

Once these parameters were determined, a method for making facial expressions as subtle as possible, but still successfully cue the target affect, was explored. For the latter, a short validation test was conducted to evaluate the following: (1) if the neutral was, indeed, formally perceived as neutral; and (2) if the shifts targeting the affects were perceived as each of their targets. For the former, it was additionally decided to create and include three versions of each facial expression at different intensity levels in the validation test, as to acquire the most subtle expression that was still above a certain recognition threshold determined by the literature.

The parameters for each emotion in its three different versions can be found in table 24 below. Still frames of all facial expressions can be found in Appendix E.

Table 23 - Parameters for lip-synch and eye movement animations.

Lip Synch Animation Parameters for Daz3D (Genesis 9 Ty Model)														
Phonemes/Frames														
Parameters	Start	G	Oh	M	I	D-B	Áh	D	Er	N-L	Áh	Sh	N	End
	27	30	34	38	40	43	48	53	55	58	62	66	70	72
Jaw Open	0%	5%	25%	0%	10%	0%	40%	5%	15%	5%	40%	2%	5%	0%
Mouth Compress	0%	20%	25%	0%	20%	0%	0%	5%	5%	5%	0%	5%	5%	0%
Mouth Funnel	0%	10%	20%	0%	10%	0%	10%	10%	10%	10%	10%	10%	10%	0%
Mouth Lower Down	0%	30%	5%	0%	30%	0%	0%	20%	20%	20%	0%	20%	20%	0%
Mouth Press	0%	0%	0%	10%	0%	0%	0%	5%	5%	5%	5%	5%	5%	0%
Mouth Pucker	0%	0%	30%	0%	0%	0%	5%	5%	5%	5%	5%	5%	5%	0%
Mouth Purse	0%	10%	10%	10%	10%	10%	0%	0%	0%	0%	0%	0%	0%	0%
Mouth Upper Up	0%	0%	10%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Eye Movement Animation Parameters for Daz3D (Genesis 9 Ty Model)													
Phonemes/Frames													
Parameters	Start	G	Oh	D-B	Áh	D	N-L	Áh	Sh	-	N	End	-
	27	30	34	43	48	53	58	62	66	68	70	72	74
Brow Up	0%	10%	0%	0%	20%	0%	0%	15%	0%	0%	0%	0%	0%
Blink	0%	0%	0%	0%	0%	0%	0%	0%	30%	90%	30%	15%	0%

Table 24 - Daz3D Parameters for each affect.

Affect Parameters for Daz3D (Genesis 9 Ty Model)			
Angry	Versions		
Parameters	Ver. 01 (Slight)	Ver. 02 (Mid)	Ver. 03 (Intense)
Brow Down	90%	100%	100%
Brow Squeeze	50%	75%	90%
Eyelid Open-Close Upper	15%	20%	30%
Mouth Press	15%	20%	30%
Mouth Frown	15%	20%	30%
Nasal Flare	15%	20%	30%

Bored	Versions		
Parameters	Ver. 01 (Slight)	Ver. 02 (Mid)	Ver. 03 (Intense)
Eyelid Open-Close Upper	-20%	-30%	-40%
Mouth Frown	0%	20%	40%

Happy	Versions		
Parameters	Ver. 01 (Slight)	Ver. 02 (Mid)	Ver. 03 (Intense)
Cheek Squint	40%	50%	60%
Eyelid Open-Close Lower	-10%	-20%	-30%
Mouth Dimple	50%	60%	70%
Mouth Smile	10%	20%	30%

Relaxed	Versions		
Parameters	Ver. 01 (Slight)	Ver. 02 (Mid)	Ver. 03 (Intense)
Mouth Dimple	10%	20%	30%
Eyelid Open-Close Upper	-10%	-20%	-30%

Sad	Versions		
Parameters	Ver. 01 (Slight)	Ver. 02 (Mid)	Ver. 03 (Intense)
Brow Inner Up	80%	90%	100%
Brow Down	80%	90%	100%
Brow Squeeze	60%	70%	80%
Mouth Frown	40%	50%	60%

Interested	Versions		
Parameters	Ver. 01 (Slight)	Ver. 02 (Mid)	Ver. 03 (Intense)
Brow Up	20%	40%	60%
Mouth Dimple	10%	20%	30%

Once the animations were finalized in Daz3D, the frames were rendered as static images and exported to Adobe Photoshop (Adobe Inc., 2023) to create the animations. This software shift was done for a few reasons: first, rendering realistic 3D animations is very computationally complex due to the surface lighting calculations required. The hardware

required for rendering animations straight from Daz3D was not available, and attempting to do so with the available machines would result in long rendering times and potential crashes midway. Rendering the individual frames as images allowed for the task to be broken down across multiple computers and, if there was a crash, the frames rendered before the error would already be properly stored, allowing for the rendering to be resumed from where it was interrupted instead of re-starting the entire process.

The second reason is the matter of frame rate. In Daz3D, the frame rate is set at the start of the project, and altering this rate through the process of animating would cause the program to adjust the timing of the frames based on the previous rate. While a static frame rate was good for most affective stimuli, the fact that the *sad* and *happy* stimuli had global tempo shifts applied across the whole utterance, changing the timing of key frames aligned to phonemes, also had to be considered. This would require calculating different lip-synch key frames for both *sad* and *happy* individually, further adding to a rather intensive process.

A simpler solution to maintain the uniformity of the animation was found through Adobe Photoshop (Adobe Inc., 2023). With the static frames, Photoshop allows one to set the frame rate of a project regardless of the starting number of frames, positioning them equidistantly. Thus, when loading the individual frames on Photoshop as an image sequence, the frame rate was set to 33 fps for *happy* (a tempo reduction of 10% from the base rate of 30 fps), and 27 fps for *sad* (a 10% tempo increase from the base rate). While this, in principle, changes the smoothness of the animation (i.e., the higher the frame rate, the more realistic and smooth the movement is), this change is small enough that it is not noticeable to the naked eye of a naïve listener, especially considering that the focus of the experiments is still the voice. Frame rate differences are usually only perceived between larger shifts (e.g., from 30 to 60 fps), and by trained eyes.

Then, the frames compiled in Photoshop were exported as animated videos, formatted for upload to YouTube for embedding on Google Forms, and tested for affect recognition before the vocal stimuli were added.

7.1.3. Visual Stimuli Validation Experiment

A short experiment to validate the visual stimuli was conducted before combining them with the vocal stimuli. The objective for this test was two-fold: (1) to evaluate if the stimuli created truly invoked the target affects when presented in isolation; and (2) to evaluate which of the three versions of each affect were the most subtle, while still standing above a certain validation threshold in terms of recognition rates.

A total of 19 animated visual stimuli (3 versions of each affect including *angry*, *happy*, *sad*, *bored*, *interested*, *relaxed* and the *neutral*) with no audio were presented to viewers in a perception test set up on Google Forms. Instructions were presented in English, and a primer set of stimuli consisting of one video presenting the neutral stimulus and a second video presenting all the facial expressions in sequence was shown to familiarize judges with the stimuli. They were also given a practice run of three stimuli before the test proper.

For the proper test, the set of 19 stimuli was presented three times. Each stimulus was presented individually (not preceded by the *neutral*) in the form of a Youtube video and, for each video, subjects were asked to judge: (1) the emotion they perceived from the facial expression, within a multiple-choice list that included *neutral* (no emotion), *angry*, *happy*, *sad*, *bored*, *relaxed*, *interested*, and a 'other' option where they were able to type in their own choice of affect; and (2) the intensity of the emotion, on a 4-point scale from neutral to slight, moderate, and intense emotion.

This test was presented to 20 subjects who, differently from previous experiments, were not required to be Irish speakers, rather being a separate group of people from varying Western European and American cultural backgrounds and fluent English speakers. The decision not to limit this test to Irish speakers was done based on the following reasons: first, to avoid exhausting the pool of Irish speakers, who would be requested to take multiple tests (divided among smaller groups to avoid too much population overlapping) following this validation experiment; and second, while some evidence of cultural specificities in facial expressions was reported in the literature review, most studies show a considerable number of cross-cultural similarities in facial emotion perception. Therefore, a separate group of subjects with relatively similar cultural backgrounds was chosen for this validation test.

Results

The results of the validation experiment, i.e., recognition rates of each visual stimulus are presented below in figure 77 and table 2. Values consist of an average of all repetitions of one stimulus presented to viewers.

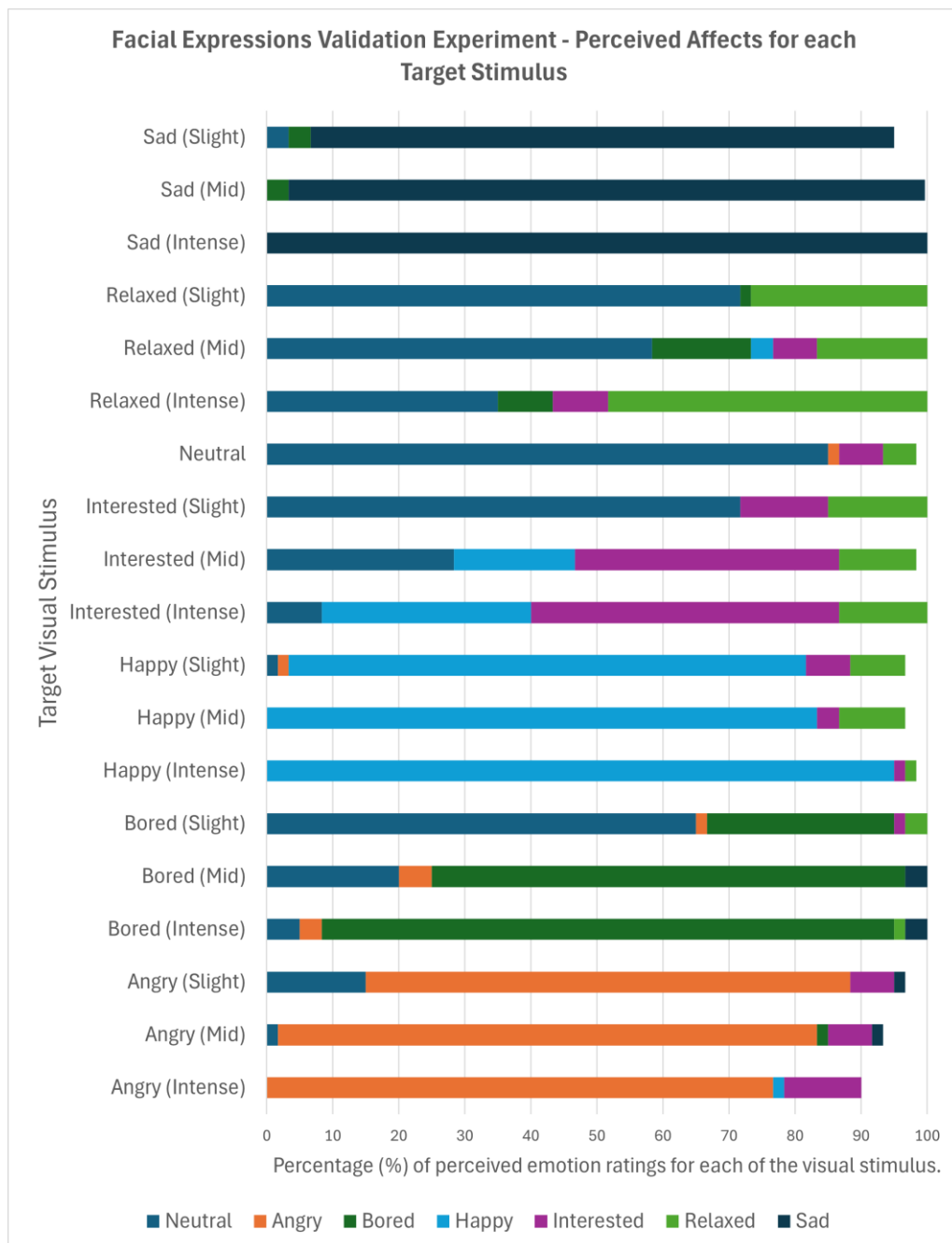


Figure 77 – Percentages of affect identification ratings (perceived affects) for each target stimulus (averaged from all repetitions). Percentage remainders for adding up to 100% represent responses for the ‘other’ option, detailed in table 26 below.

Table 25 - Visual Stimuli Experiment Results as percentages (%) (averaged from repetitions and rounded to the nearest integers) of affect perception rates for each target visual stimulus.

Stimuli	Perceived Affect (% of responses)						
	<i>Neutral</i>	<i>Angry</i>	<i>Bored</i>	<i>Happy</i>	<i>Interested</i>	<i>Relaxed</i>	<i>Sad</i>
Angry (Intense)	0	77	0	2	12	0	0
Angry (Mid)	2	82	2	0	7	0	2
Angry (Slight)	15	73	0	0	7	0	2
Bored (Intense)	5	3	87	0	0	2	3
Bored (Mid)	20	5	72	0	0	0	3
Bored (Slight)	65	2	28	0	2	3	0
Happy (Intense)	0	0	0	95	2	2	0
Happy (Mid)	0	0	0	83	3	10	0
Happy (Slight)	2	2	0	78	7	8	0
Interested (Intense)	8	0	0	32	47	13	0
Interested (Mid)	28	0	0	18	40	12	0
Interested (Slight)	72	0	0	0	13	15	0
Neutral	85	2	0	0	7	5	0
Relaxed (Intense)	35	0	8	0	8	48	0
Relaxed (Mid)	58	0	15	3	7	17	0
Relaxed (Slight)	72	0	2	0	0	27	0
Sad (Intense)	0	0	0	0	0	0	100
Sad (Mid)	0	0	3	0	0	0	96
Sad (Slight)	3	0	3	0	0	0	88

From this, one can note that all versions of *angry*, *happy* and *sad* achieved very high target-affect recognition ratings. *Happy* and *sad* had the expected outcome – the more intense the emotion portrayal, the higher its recognition. For *angry*, the middle ground arose as the most effective. For *bored*, the intense and mid versions also achieved very high recognition ratings, the former even higher than *angry*. These stimuli were all rarely perceived as other affects, being uniquely recognized within their activation categories even though they were created with subtlety in mind from the start, and even for the more

intense versions. The slight *bored*, however, was more often perceived as the *neutral* than as its target affect.

Interested and *relaxed* were the affects that received poorer target-affect recognition ratings. Both intense versions only scored barely above the 45% average; *interested*-mid achieved 40% recognition, but its subtlest version is much more often perceived as *neutral*. This effect is even more obvious in *relaxed* – as the affect portrayal gets subtler, its rating as *neutral* increases, being even higher than ratings for the *relaxed*-target. Therefore, only the most intense versions of these affects were successful at conveying the target affect above chance, and above *neutral* ratings. Additionally, two subjects made use of the ‘other’ option, perceiving the following affects from the faces (sometimes varying across repetitions) as shown in table 26:

Table 26 - Additional affects associated with some of the visual stimuli.

Target Affect	Perceived Affects
Angry (Intense)	Scared, serious, surprised, afraid.
Angry (Mid)	Afraid, intense, serious.
Angry (Slight)	Afraid, concerned.
Happy (Intense)	Playful.
Happy (Mid)	Nefarious, Playful (happy).
Interested (Mid)	Happy or interested.
Neutral	Serious.
Sad (Slight)	Concerned, worried.

Given these results, the choice for the final set of visual stimuli to be used for the larger perception test in combination with the vocal stimuli was done as follows. Scherer (2003) mentions an average recognition rate of around 75% for emotions on the face. However, he also states that some emotions are more easily recognized than others, as this average includes emotions such as disgust which achieved a recognition rate of only 31% (i.e., barely above chance). Similar results can be found in Keltner, Sauter, et al. (2019), reporting a cross-cultural study. Boredom, anger, and happiness all score above 70%, but sadness performed worse in their experiment (just above 50%) than what the results above show. Accordingly, their results for *interested* mirror what is seen here (about 45% compared to 48.3% for our *intense interested*).

For *relaxed*, results across the literature are not as straight forward to interpret, particularly since not many studies explore this or similar affects. As these results show, *relaxed* was most often perceived as *neutral* in its slighter variations, with only the most intense version achieving above average recognition as the target. In contrast, Keltner, Sauter, et al. (2019) reports a 65% recognition rate from a previous experiment for what they call ‘*contentment*’, namely the closest emotion in their set to the *relaxed* affect presented here (which also utilizes the same action units). They, however, did not include a neutral face in the set, which may have led to less ambiguity for test subjects as the other affects presented were more differentiated from it. Additionally, a validation study by Tottenham et al. (2009) evaluated the recognition of ‘calmness’ in the face. They also found a higher rate than the results presented here, but, on the other hand, those are also merged with their own neutral, making it hard to judge how these stimuli would perform on their own but in the presence of a separate individual neutral stimulus.

Therefore, to maintain the desired subtlety, a 70% threshold was established for the more successful emotions to allow for a slight ambiguity margin in comparison to the 75% average from Scherer (2003). The more subtle versions of each affect-targeting facial expression were chosen if they were, on average, recognized above this threshold. As for *interested* and *relaxed*, it was decided to use the more intense versions of each of these affects, as both achieved a target recognition rate above 45%. The final choice of stimuli to be presented with the vocal stimuli are therefore as follows: ***angry (slight), bored (mid), happy (slight), interested (intense), relaxed (intense), and sad (slight)***. Neutral was also maintained as it achieved a high recognition rate.

Once these stimuli were validated and the choice for the variations made, the animations were combined with the vocal stimuli in Adobe Photoshop (Adobe Inc., 2023), with careful attention paid to the lip-synch timing. The first set of stimuli were combined congruently between target affects (i.e., the *happy* visual with the *happy* vocal stimulus); then, incongruent stimuli were created, with attention to frame rate/timing shifts. These incongruent combinations are further detailed in each of the experiments detailed below.

7.2. Experiment 3 – Visual Context with Cross-Activation Congruence

The first experiment conducted with the combined vocal and visual stimuli is described in this subsection. The chosen and validated visual stimuli detailed in the previous section were combined with the vocal stimuli. More specifically, the set of vocal stimuli used were the combined stimuli of each affect, using the ABAIR synthesizer neutral as a baseline with the additional scaling applied to EE and RD, and created with the final version of the VSG (Z. Wang & Gobl, 2023b) as described.

The decision to use only the combined stimuli rather than all versions of the vocal stimuli (*global*, *local*, and *combined*) was made for two reasons. First, using all the versions of the stimuli would add too many variables to the test, considering all the different combinations of visual and vocal stimuli already planned. It would therefore have been difficult to judge which variable was responsible for the affect recognition ratings, given the presence of the visual stimuli and the differences in scaling from the experiment described in chapter 6. Additionally, it would also result in a very large number of stimuli in the listening test, which would be potentially exhausting for the informants. Secondly, the combined stimuli were the closest to the vocal production data in terms of parameter levels. Even though additive effects of global and local stimuli on the target recognition rate of all affects were not found, the combined stimuli were the most realistic in terms of stylization, best matching the contours of the production data on average. Since this experiment did not aim to test the effects of global and local shifts in the voice, it was decided to uniformly use the combined versions of all vocal affects.

This experiment therefore focused on the effects of the visual stimuli on the affect recognition of the vocal stimuli, as the latter were presented to subjects (1) on their own, (2) combined with emotionally congruent visual stimuli, and (3) combined with emotionally incongruent visual stimuli. To avoid the large number of combinations that would result from combining all affects across each other, it was decided to split these into two tests. The test presented in this subsection explores incongruent stimuli that consist of combinations across activation categories – i.e., what would be considered ‘opposite’ affects, as presented in the scales used for the test in chapter 6 and again in the experiments described here. A summary of the stimuli combinations can be found in table 27 below.

Table 27 - Vocal and visual stimuli combinations for the perception test.

Vocal Stimulus	Visual Stimulus		
	None	Congruent	Incongruent
Angry	None	Angry	Relaxed
Bored	None	Bored	Interested
Happy	None	Happy	Sad
Relaxed	None	Relaxed	Angry
Sad	None	Sad	Happy
Neutral	None	Neutral	N/A

As there was no *interested* vocal stimulus, the visual stimulus was only used as the incongruent combination to *bored*. The 17 stimuli were then produced in video format (the voice-only stimuli presented a black background with the words ‘éist leis’ (*‘listen to him’*) in white lettering), uploaded to YouTube and embedded into the listening test.

7.2.1. Perception Test Design and Subjects

The 17 stimuli were presented to listeners in a perception test set up on Google Forms, with questions in the same format as experiment 2, illustrated by figure 70 (for task 1). They were given a set of simple instructions in Irish¹², including listening to the stimuli with headphones in a quiet environment. They were then presented with a primer set of stimuli, consisting of one version of the neutral vocal stimulus separately, and a sequence of all the vocal-only stimuli (without the visuals) to familiarize themselves with them, and have the chance to adjust their system volumes based on the lower intensity of the low activation affects.

Like experiment 2 in chapter 6, this test was divided into three scale-based tasks. A practice run with 6 stimuli (2 voice only, 2 congruent and 2 incongruent) was presented before the first task to familiarize subjects with the test, in the format of the first task.

Each task included 17 pairs of stimuli: the neutral stimulus, followed by one of the 17 stimuli (including neutral, repeated), presented in a randomized sequence. The preceding neutral was presented in voice-only form if followed by another voice-only stimulus, and with the congruent neutral visual stimulus if followed by one of the visual and

¹² Translations from English to Irish were done by Prof. Ailbhe Ní Chasaide.

voice combination stimuli. The tasks were presented in the same format as experiment 2: for each task, listeners were asked to judge the emotion of the second utterance of the pair on a 7-point scale of intensity between -3 (low activation) and +3 (high activation), each polarity representing one of two opposing affects. Task 1 involved the *sad* vs. *happy* affects; task 2 asked them to choose between *bored* vs. *interested*; and task 3 between the *relaxed* vs. *angry* affects. If listeners perceived one of the two affects in the scale, they were asked to make their judgement based on the numerical scale, with -3 being a strong expression of the low activation affect in the scale and +3 being a strong expression of the high activation affect. The ± 2 represented a moderate expression of that emotion, and ± 1 represented a slight expression. Listeners were asked to choose the midpoint 0 (zero) whenever no affect was perceived (no emotion/neutral), or when they perceived an affect from the stimulus that was not on any end of the scale presented in that subtest. Again, the *interested* option was only included as a counterpart to *bored*, but no vocal stimuli targeting *interested* were included in the set.

The perception test was presented to 31 participants with varying levels of proficiency in Irish. Because a practice run was included before the test, no responses to the actual sets of 17 stimuli were discarded.

7.2.2. Statistical Analysis and Results

Statistical approaches and results of this test are presented below. For the statistical analysis, mixed-effects ordinal logistic regression models were used to measure the relationships between the intensity scores and target affects/visual stimulus combination types through R (R Core Team, 2021). The same model from experiment 2 was used, but adapted to the new variables. Three separate models were fitted, one for each affect dimension present in the perception test: *sad-happy*, *bored-interested*, and *relaxed-angry*. The 7-point scale (between -3 and +3) used for participant ratings was treated as an ordinal dependent variable, with target affect and visual stimulus combination type being treated as independent variables, which were tested separately. The model also included the interactions between target affect and combination type, with random intercepts for listener ID to account for individual differences. Holm-Bonferroni adjustments (Holm, 1979) were also applied to account for multiple comparisons. Confidence intervals were calculated using standard errors from the model output, and *p* values below .05 after the

Holm-Bonferroni adjustment were considered statistically significant. Reference categories used for the analysis were ‘neutral’ for the target affect, and voice-only (‘none’) for the visual stimuli combination type. Model coefficients represent the log-odds of stimuli moving to a higher rating category on the ordinal scale (high activation affect categories), with negative values indicating a move towards the negative end of the scale (low activation categories).

Like in previous experiments, an overall model could not be fitted across all scales at once due to the nature of the experiment (i.e., only two of the target affects were available as perceived affects for each task), which caused numerical issues and quasi-complete separation of data during modelling. The per-scale model was therefore considered the most suitable and robust for the analysis presented. Since the experiment was conducted with a single measure to capture both affect identification and intensity rate, the model was written so that all significant values shown in the tables and figures indicate significance on both of those dimensions. Further clarification is given below. Results are presented in figures 78-83, and tables 28-30. As before, absent interactions were dropped due to rank deficiency that could cause coefficient estimation issues.

Figure 78 presents a heatmap confusion matrix showing listener’s judgements of affect (perceived affect) per scale, based on each target affect and visual stimulus combination type, based only on a binary identification of affect (e.g., for the *sad-happy* scale, any positive values are counted as one instance of *happy* as the perceived affect, and negative values count as an instance of *sad*, and so on). Figure 79 presents an alternative representation of affect recognition rates for each affect-targeting triad of combinations, also based on binary preprocessing. Figures 80-82 show the results of the ordinal regression analysis as confidence intervals over the coefficient estimates for the effects of target affect, combination type and their interactions for each test scale. Finally, figure 83 shows the confidence intervals for each combination triad of affect-targeting stimuli on each scale, plotting the confidence intervals levelled to the intensity scale of the perception test.

PERCEIVED AFFECT RATINGS (TARGET VOCAL AFFECT + VISUAL COMBINATION TYPE)

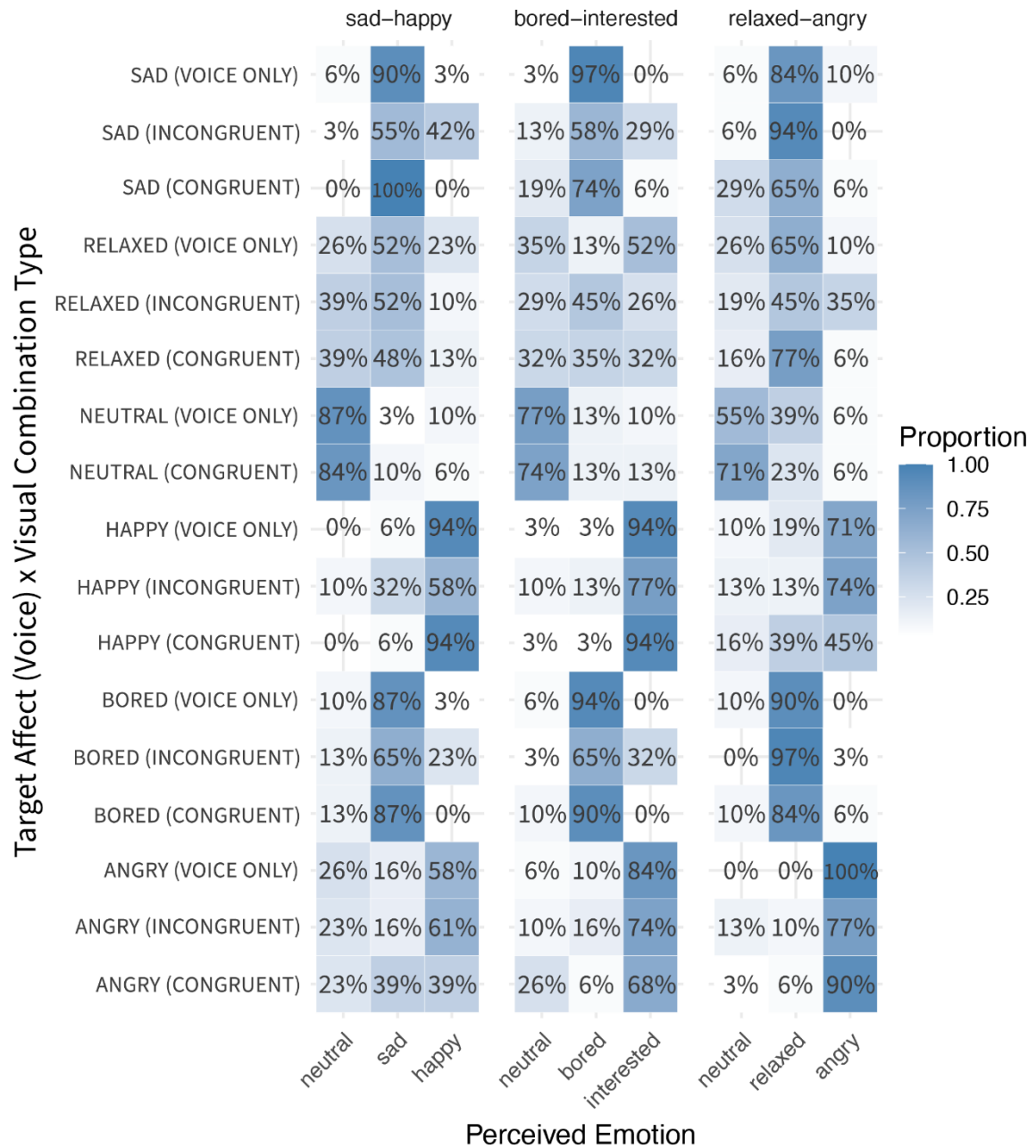


Figure 78 - Affect identification heatmap matrix showing the percentages of affect identification rates (perceived affect) based on target affect and visual combination types, and on binary identification of affect (i.e., considering only categorical identification on each end of the test scales, without the effects of intensity scores).

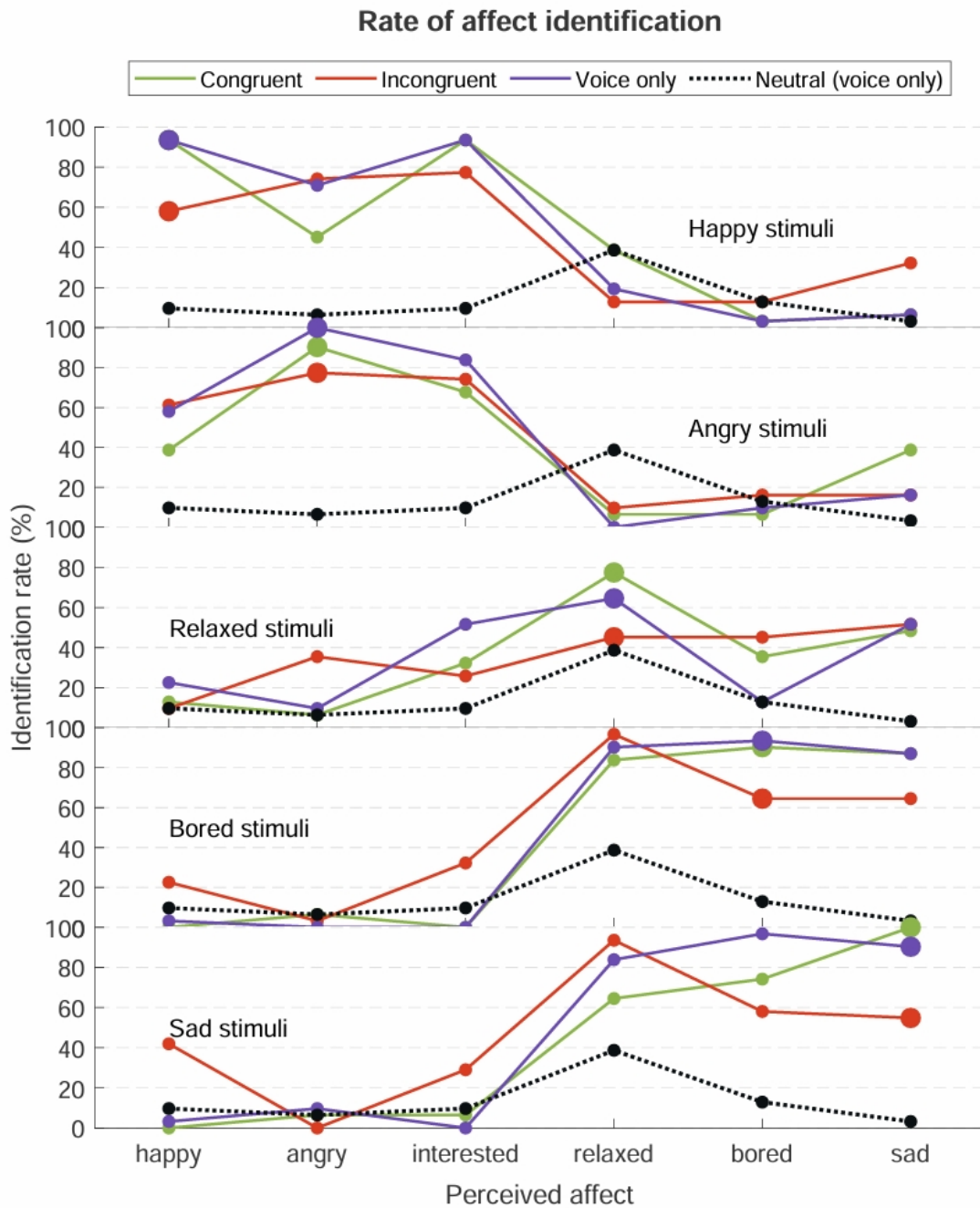


Figure 79 - Alternative representation of affect identification rates for each combination triad of affect-targeting stimuli. Titles by each section of the image indicate the stimuli's target affect. Affect listings on the horizontal axis represent perceived affect from listeners, and the vertical axis represents percentages of identification rates. Identification rates are based on binary identification of affect (i.e., considering only categorical identification on each end of the test scales, without the effects of intensity scores). Neutral is shown as the dotted black line for comparison. The larger dots show the values of the target affect for each panel.

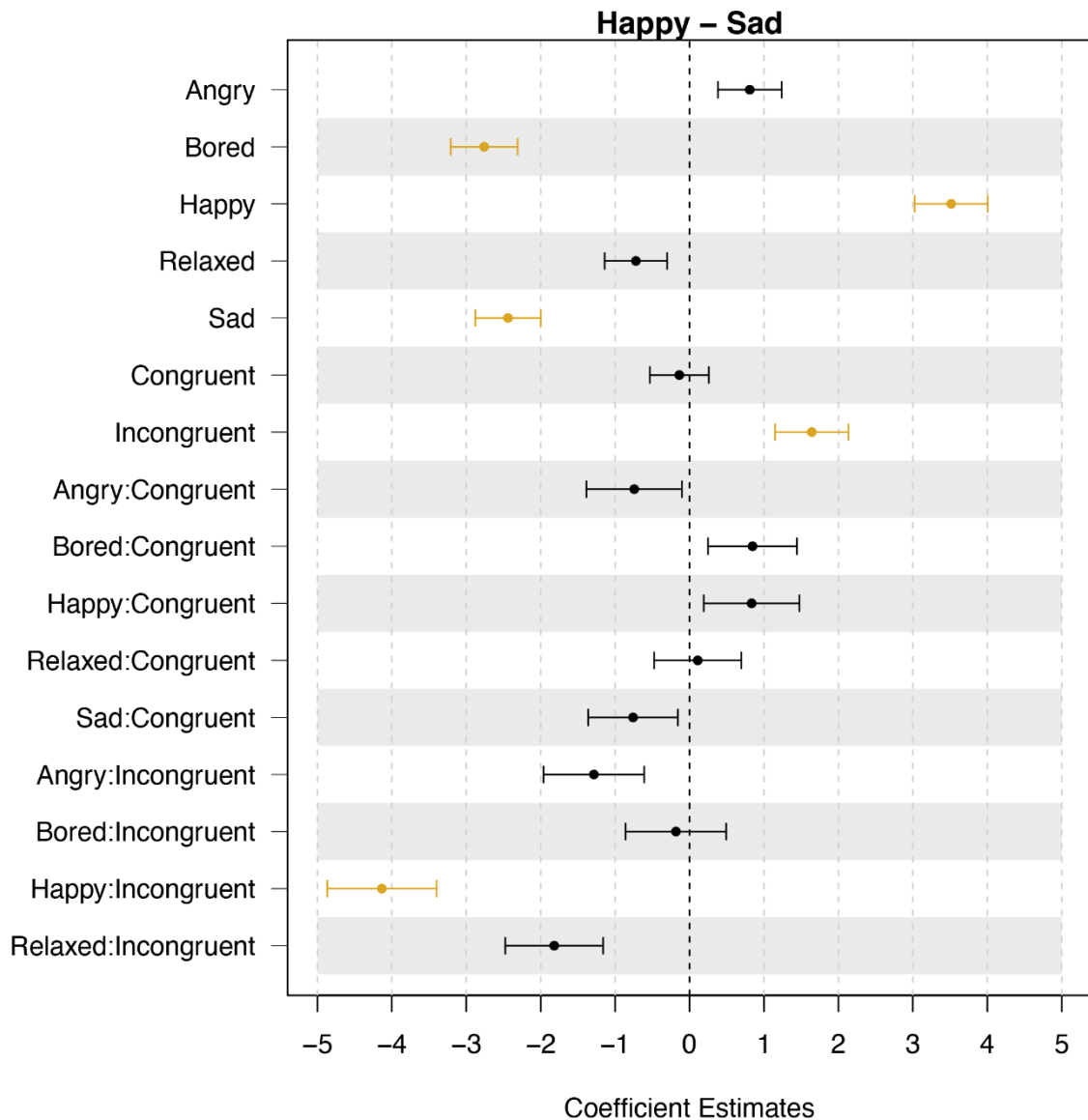


Figure 80 – Coefficient estimates for target affect, combination type and interaction effects from the mixed-effects ordinal regression analysis results for the *sad-happy* scale. Yellow lines indicate Holm-Bonferroni adjusted significances. Sidebars indicate Standard Error to each side. Bars completely above 0 on the horizontal axis indicates positive correlation, bars intersecting 0 indicate no correlation, and bars below 0 indicate negative correlation (see table 28 for numerical values).

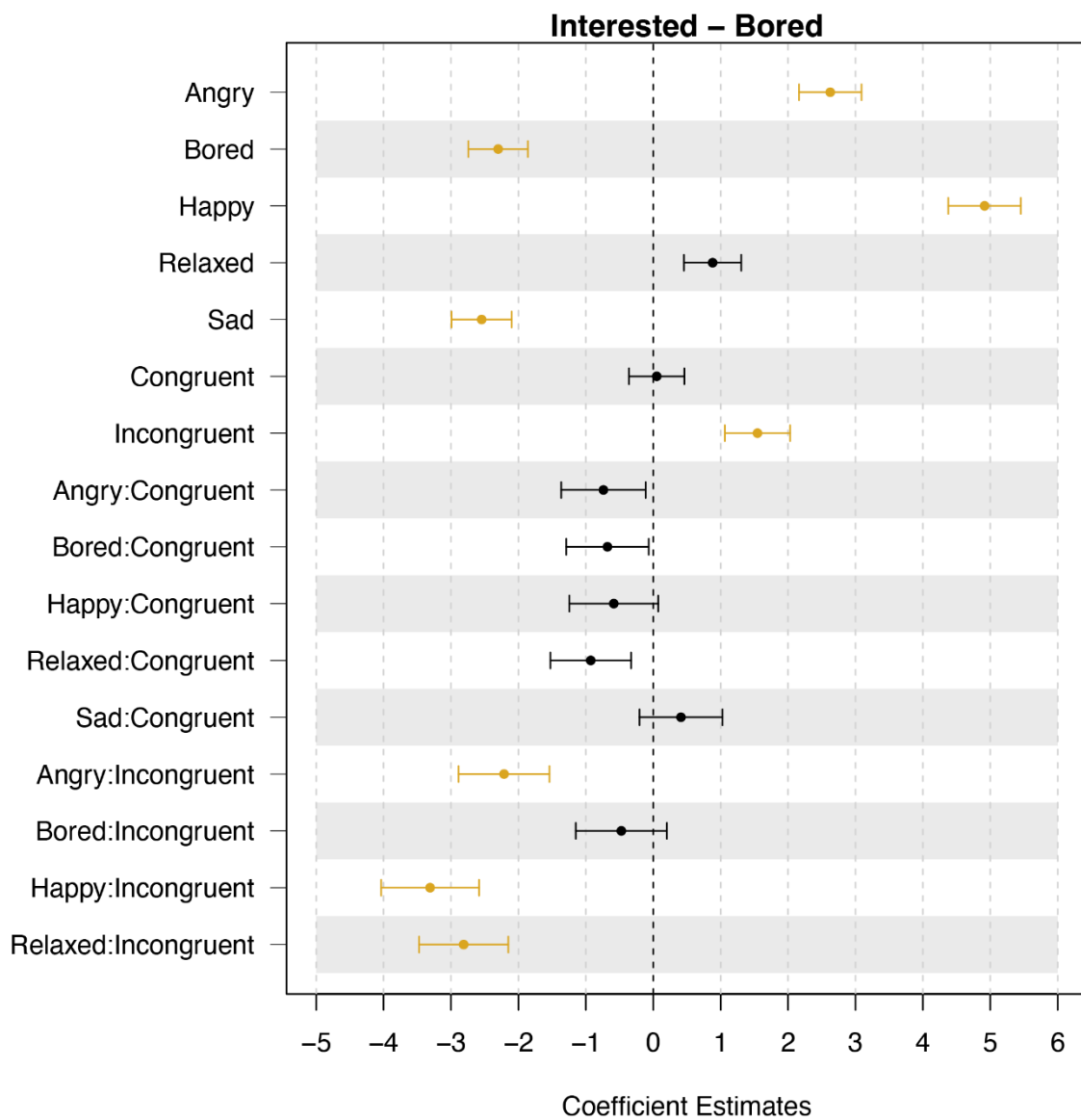


Figure 81 – Coefficient estimates for target affect, combination type and interaction effects from the mixed-effects ordinal regression analysis results for the *bored-interested* scale. Yellow lines indicate Holm-Bonferroni adjusted significances. Sidebars indicate Standard Error to each side. Bars completely above 0 on the horizontal axis indicates positive correlation, bars intersecting 0 indicate no correlation, and bars below 0 indicate negative correlation (see table 29 for numerical values).

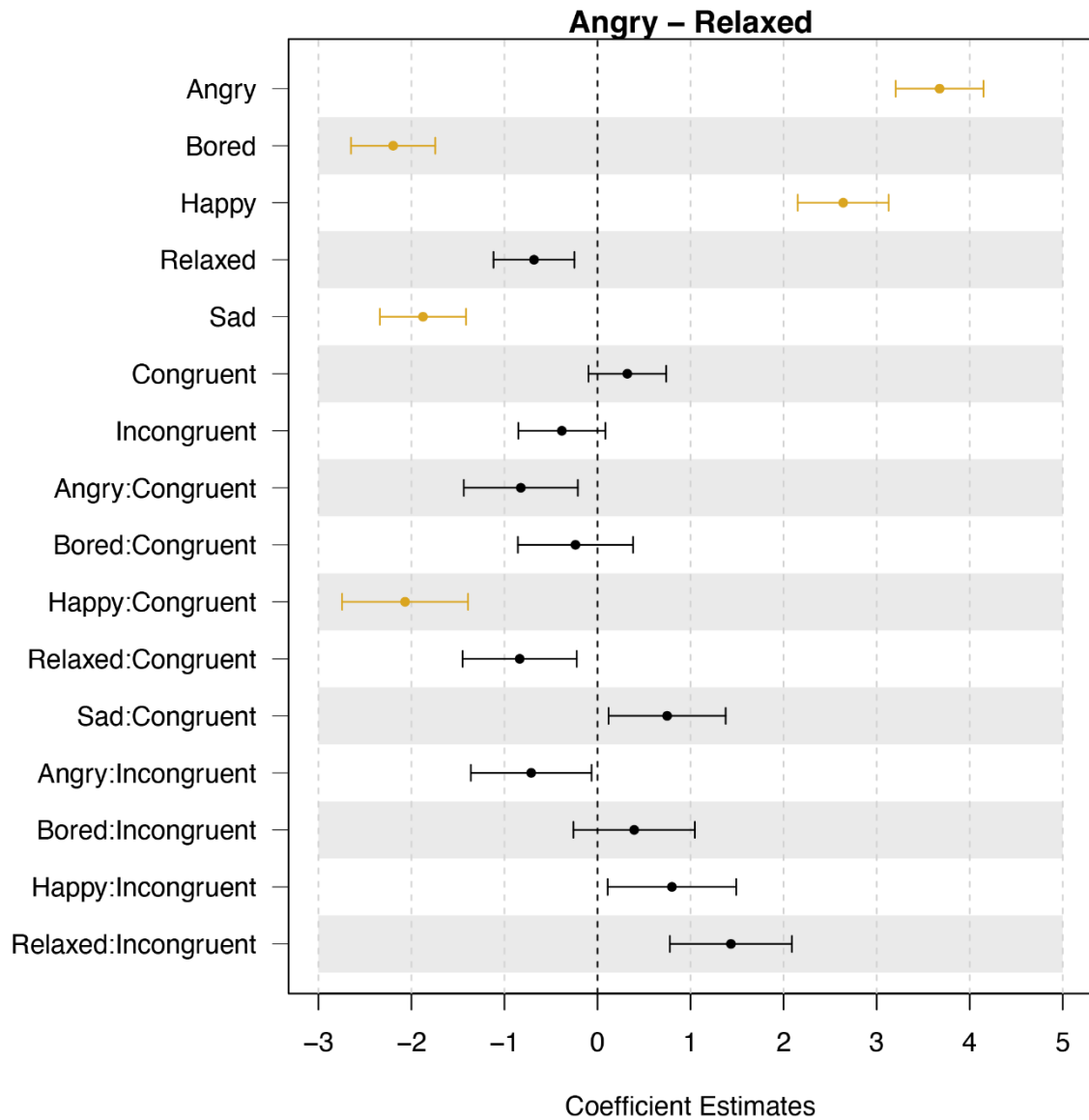


Figure 82 – Coefficient estimates for target affect, combination type and interaction effects from the mixed-effects ordinal regression analysis results for the *relaxed-angry* scale. Yellow lines indicate Holm-Bonferroni adjusted significances. Sidebars indicate Standard Error to each side. Bars completely above 0 on the horizontal axis indicates positive correlation, bars intersecting 0 indicate no correlation, and bars below 0 indicate negative correlation (see table 30 for numerical values).

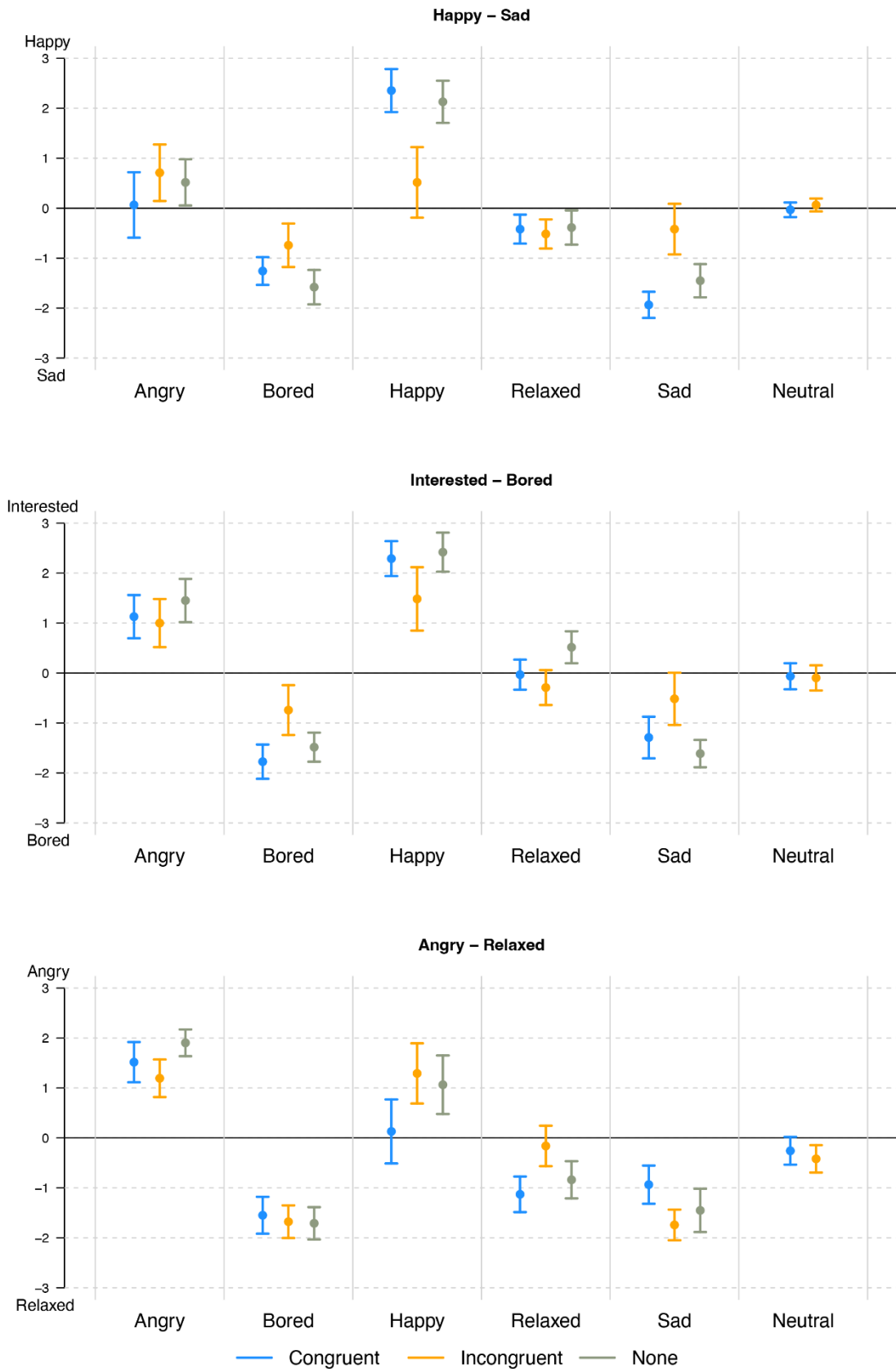


Figure 83 – Intensity score confidence Intervals for all target affects and triads of visual combinations across all scales. Error bars indicate 95% confidence intervals (± 2 standard errors) around each mean.

Table 28 - Ordinal logistic regression results for the *sad-happy* scale with Holm-Bonferroni adjustments, showing main effects of target affect, visual stimulus combination type and interactions on listeners' perception of affect intensity. Positive estimates indicate positive correlation/above-chance performance relative to the reference categories; negative estimates indicate negative correlation/below-chance performance. Values close to 0 indicate no significant effect. Reference categories are *neutral* (target emotion) and *voice-only* (combination type). Absent interactions were dropped due to rank deficiency from showing patterns too similar to the reference levels. Significance levels: * $p < .05$, ** $p < .01$, *** $p < .001$.

**Target Emotion × Visual Stimulus Combination Type Effects on Intensity Ratings
(Sad-Happy Scale)**

Predictor	Estimate (β)	SE	Z-value	P-value	P-value (Holm-Adjusted)
Main Effects					
Target Affect: Angry	0.81	0.43	1.89	.06	.58
Target Affect: Happy	3.52	0.49	7.16	<.001***	<.001***
Target Affect: Bored	-2.76	0.45	-6.14	<.001***	<.001***
Target Affect: Sad	-2.44	0.44	-5.55	<.001***	<.001***
Target Affect: Relaxed	-0.72	0.42	-1.72	.09	.69
Visuals: Congruent	-0.14	0.40	-0.35	.73	1.00
Visuals: Incongruent	1.64	0.49	3.34	<.001***	.01*
Interactions					
Angry vs. Congruent	-0.74	0.64	-1.16	.25	1.00
Bored vs. Congruent	0.85	0.60	1.42	.16	1.00
Happy vs. Congruent	0.83	0.64	1.30	.19	1.00
Relaxed vs. Congruent	0.11	0.59	0.19	.85	1.00
Sad vs. Congruent	-0.76	0.60	-1.26	.21	1.00
Angry vs. Incongruent	-1.29	0.68	-1.90	.06	.58
Bored vs. Incongruent	-0.18	0.68	-0.27	.79	1.00
Happy vs. Incongruent	-4.13	0.73	-5.63	<.001***	<.001***
Relaxed vs. Incongruent	-1.82	0.66	-2.77	.01**	.06

Table 29 - Ordinal logistic regression results for the *bored-interested* scale with Holm-Bonferroni adjustments, showing main effects of target affect, visual stimulus combination type and interactions on listeners' perception of affect intensity. Positive estimates indicate positive correlation/above-chance performance relative to the reference categories; negative estimates indicate negative correlation/below-chance performance. Values close to 0 indicate no significant effect. Reference categories are *neutral* (target emotion) and *voice-only* (combination type). Absent interactions were dropped due to rank deficiency from showing patterns too similar to the reference levels. Significance levels: * $p < .05$, ** $p < .01$, *** $p < .001$.

***Target Emotion × Visual Stimulus Combination Type Effects on Intensity Ratings
(Bored-Interested Scale)***

<i>Predictor</i>	<i>Estimate (β)</i>	<i>SE</i>	<i>Z-value</i>	<i>P-value</i>	<i>P-value (Holm-Adjusted)</i>
<i>Main Effects</i>					
Target Affect: Angry	2.63	0.46	5.65	<.001***	<.001***
Target Affect: Happy	4.92	0.54	9.14	<.001***	<.001***
Target Affect: Bored	-2.30	0.44	-5.21	<.001***	<.001***
Target Affect: Sad	-2.55	0.45	-5.70	<.001***	<.001***
Target Affect: Relaxed	0.88	0.42	2.08	.04*	.30
Visuals: Congruent	0.05	0.41	0.12	.90	1.00
Visuals: Incongruent	1.55	0.48	3.19	.001**	.01*
<i>Interactions</i>					
Angry vs. Congruent	-0.74	0.63	-1.18	.24	1.00
Bored vs. Congruent	-0.68	0.61	-1.11	.27	1.00
Happy vs. Congruent	-0.59	0.66	-0.89	.38	1.00
Relaxed vs. Congruent	-0.93	0.60	-1.55	.12	.85
Sad vs. Congruent	0.41	0.61	0.67	.50	1.00
Angry vs. Incongruent	-2.21	0.67	-3.28	.001**	.01*
Bored vs. Incongruent	-0.47	0.68	-0.70	.48	1.00
Happy vs. Incongruent	-3.31	0.73	-4.56	<.001***	<.001***
Relaxed vs. Incongruent	-2.81	0.66	-4.25	<.001***	<.001***

Table 30 - Ordinal logistic regression results for the *relaxed-angry* scale with Holm-Bonferroni adjustments, showing main effects of target affect, visual stimulus combination type and interactions on listeners' perception of affect intensity. Positive estimates indicate positive correlation/above-chance performance relative to the reference categories; negative estimates indicate negative correlation/below-chance performance. Values close to 0 indicate no significant effect. Reference categories are *neutral* (target emotion) and *voice-only* (combination type). Absent interactions were dropped due to rank deficiency from showing patterns too similar to the reference levels. Significance levels: * $p < .05$, ** $p < .01$, *** $p < .001$.

**Target Emotion × Visual Stimulus Combination Type Effects on Intensity Ratings
(Relaxed-Angry Scale)**

Predictor	Estimate (β)	SE	Z-value	P-value	P-value (Holm-Adjusted)
Main Effects					
Target Affect: Angry	3.68	0.47	7.78	<.001***	<.001***
Target Affect: Happy	2.64	0.49	5.41	<.001***	<.001***
Target Affect: Bored	-2.20	0.45	-4.85	<.001***	<.001***
Target Affect: Sad	-1.88	0.46	-4.06	<.001***	<.001***
Target Affect: Relaxed	-0.68	0.44	-1.57	.12	1.00
Visuals: Congruent	0.32	0.42	0.77	.44	1.00
Visuals: Incongruent	-0.38	0.47	-0.82	.41	1.00
Interactions					
Angry vs. Congruent	-0.82	0.61	-1.35	.18	1.00
Bored vs. Congruent	-0.24	0.62	-0.38	.70	1.00
Happy vs. Congruent	-2.07	0.68	-3.06	.002**	.03*
Relaxed vs. Congruent	-0.84	0.61	-1.37	.17	1.00
Sad vs. Congruent	0.75	0.63	1.19	.24	1.00
Angry vs. Incongruent	-0.71	0.65	-1.10	.27	1.00
Bored vs. Incongruent	0.39	0.65	0.60	.55	1.00
Happy vs. Incongruent	0.80	0.69	1.16	.25	1.00
Relaxed vs. Incongruent	1.43	0.66	2.19	.03*	.32

The results of the separate ordinal regression models fitted on a per-scale basis are presented here. Each model examined the effects of the target affects and visual stimuli combination type (voice-only, congruent and incongruent) on the perception of affect by participants. To reiterate, reference categories used in this analysis were neutral for target affect, and voice-only ('none') for the visual combination types.

The initial hypothesis was that the congruent visual stimuli would strengthen the perception and increase the target identification rates of all affects compared to the voice-only stimuli, while the incongruent visual stimuli would hinder this identification to produce a lower rating than for the voice-only stimuli. It was also hypothesized that the addition of the visual stimuli would help in the unique signaling of these affects to a greater degree than seen in the voice-only experiments in previous sections. Once again, results appear to be mixed and affect-dependent in some cases, as detailed below.

In terms of main effects (tables 28-30, figures 80-82), consistent patterns emerged in how different target affects were rated relative to the neutral stimuli across all three scales. Once again, *happy* consistently received the highest ratings across all scales (*sad-happy*: $\beta = 3.52, p < .001$; *bored-interested*: $\beta = 4.92, p < .001$; *relaxed-angry*: $\beta = 2.64, p < .001$), being only lower than *angry* in the *relaxed-angry* scale. This indicates that the *happy*-targeting stimuli were all consistently perceived as significantly more intense towards the high activation sides of the scales. The *angry*-targeting stimuli also received significantly high activation ratings on the *bored-interested* scale ($\beta = 2.63, p < .001$) and *relaxed-angry* scale ($\beta = 3.68, p < .001$), but not on the *sad-happy* scale ($\beta = 0.81, p = .58$), suggesting that, this time and in terms of main effects, the *angry*-targeting stimuli did not conjure the *happy* affect as much as previously seen.

Low activation affects (*bored* and *sad*) consistently received significant ratings towards the low activation ends of the scales, with low intensity ratings. The *bored*-targeting stimuli were significantly rated as more *sad* ($\beta = -2.76, p < .001$), less *interested* ($\beta = -2.30, p < .001$), and more *relaxed* ($\beta = -2.20, p < .001$) on each of these perceived affect scales than the neutral reference. Similarly, *sad*-targeting stimuli were also rated as significantly more *sad* ($\beta = -2.44, p < .001$), less *interested* ($\beta = -2.55, p < .001$), and more *relaxed* ($\beta = -1.88, p < .001$). The *relaxed*-targeting stimuli, on the other hand, showed more variable patterns, being rated as significantly leaning more towards the low activation side of the *sad-happy* scale ($\beta = -0.72, p = .69$, non-significant after correction) but not showing significant effects on other scales after multiple comparisons.

Regarding visual combination types, the incongruent combinations consistently produced significant main effects across the *sad-happy* ($\beta = 1.64, p = .01$) and *bored-interested* ($\beta = 1.55, p = .02$) scales, suggesting that the mismatch in the vocal and visual

channels generally shifted ratings towards the high activation affects. The congruent combinations, however, showed no significant main effects across any scale, indicating that these matching visual and vocal stimuli combinations did not systematically alter ratings in relation to the voice-only stimuli, going against our initial hypothesis.

However, as seen in the previous experiment (chapter 6), main effects can be misleading; therefore, an analysis on the interaction effects, seen in the bottom halves of tables 28-30 and figures 80-83 are now presented.

For the *happy*-targeting stimuli, the incongruent *happy* showed a strong negative effect in the *sad-happy* ($\beta = -4.13, p < .001$) and *bored-interested* ($\beta = -3.31, p < .001$) scales, suggesting that, when the *sad* visual was added to the *happy* voice, the typical strong positive effect that the *happy* congruent and voice-only stimuli have towards the high activation ends of these scales was substantially reduced, diminishing perceived happiness and interest. This is reflected in figures 78-79, as binary identification rates for *happy* drop significantly between the voice-only and congruent *happy* (94% in the *sad-happy* scale) to 58%. The *happy* congruent, on the other hand, showed a significant negative interaction on the *relaxed-angry* scale ($\beta = -2.07, p = .03$, figure 82), suggesting that adding the congruent visual pulled this stimulus towards the low activation category, but also towards an affect of the same positive valence as *happy* (in contrast to *angry*) relative to the voice-only stimulus. Indeed, percentage ratings in figure 78 for the *relaxed-angry* scale are much closer for the *happy* incongruent. No significant differences were found between the *happy* congruent and voice-only stimuli for the other scales, indicating very similar ratings both in terms of affect identification and perceived intensity, which is well illustrated by the confidence intervals in figure 83.

The *angry*-targeting stimuli had, again, less significant effects than the *happy*-targeting stimuli. The incongruent *angry* showed a significant negative interaction on the *bored-interested* scale ($\beta = -2.81, p < .001$), indicating that the addition of the *relaxed* visual stimulus pulled the affect perception from the voice away from the high activation end of the scale (shown to a smaller degree in figure 78 as a 10% decrease in its rating as *interested*). The congruent *angry*, however, showed no significance across any scales, but actually performed a little under the voice-only stimuli, which might indicate that, strangely, the addition of the congruent visual stimulus slightly hindered the perception of high activation

from the voice (also seen in the binary identification rates in figures 78-79 and in the intensity confidence intervals in figure 83). Because there was no significant effect of the congruent combination on the *sad-happy* scale, it appears that, again, this stimulus was not well differentiated from *happy*.

As for *bored* and *sad*, while these two target affects did show significant main effects across all scales, no significant interactions with visual combination types emerged. This may indicate that these affects maintain similar patterns that don't deviate much from the voice-only reference level, regardless of the type of visual added, being kept consistently on the low activation, low perceived intensity side of the scales (as seen in figure 83). The same can be said for the *sad*-targeting stimuli, further continuing the trend seen in previous experiments presented in this study that significant differentiation between these two affects is very difficult to achieve, even with the addition of the visual channel. However, figures 78 and 79 do seem to indicate a small positive effect of the congruent *sad* stimulus on improving its perception as *sad* and lowering its perception as *bored* and *relaxed* compared to the voice-only stimulus, even if this differentiation did not emerge in the statistical analysis as significant. This may indicate that, for the *sad* affect, the congruent face is a slightly stronger cue than the current voice stimuli in the differentiation between these two affects than it is for the *bored* affect, and more emotionally intense visual cues might contribute further.

As before, the incongruent *sad* interaction was dropped, possibly because its patterns were too similar to *bored* for the model. Differences between main effects and interactions seen here, while less drastic than those that emerged in experiment 2, can still be explained in similar terms – the main effects show significant effects across averages, while the interaction effects show the specific cases where these general patterns don't apply.

As for the *relaxed-targeting* stimuli, the incongruent combination with the *angry* visual stimulus seemed to have caused a significant negative interaction on the *bored-interested* scale ($\beta = -2.81$, $p < .001$, figure 81), the face therefore drawing the stimulus towards the same activation category, but towards the opposite valence than *relaxed* compared to the voice-only stimuli. Again, this pattern is repeated in figures 78-79, but the main indication of this significance in figure 83 is that the incongruent stimulus is on the

opposite side of the scale from the voice-only stimulus, even though intensity ratings are similarly low. As also seen in figures 78-79 for the *relaxed-angry* scale, the congruent *relaxed* stimulus does seem to achieve some degree of improvement in its perception of *relaxed*, but this was not confirmed by the analysis.

In terms of affect differentiation, the activation categories seem to be well-defined, with only the voice-only *relaxed* crossing towards *interested* (figure 83), which is an affect of the same valence. The high activation affects present high binary recognition rates within their own categories, but the effects of the visual stimuli vary. The significance found for the *happy* congruent in the *relaxed-angry* scale indicate that the addition of the congruent visual helped to significantly differentiate this affect from *angry* relative to the voice-only stimulus, indicating that both the voice-only and congruent stimuli achieved a successful affect-specific signaling of happiness between those two affects. However, both the *happy* and *angry* voice-only and congruent stimuli were very often perceived as *interested* and not well differentiated, mirroring previous results. According to figure 78, the congruent and voice-only *happy* showed no differentiation between being perceived as *happy* or as *interested*, and *angry* only had a small but non-significant difference between its perception as *angry* and *interested*. The addition of the congruent *angry* visual did not seem to significantly improve that differentiation, either.

For the low activation affects, no additional significant differentiation was achieved through the addition of the congruent visuals, even though a small effect is apparent in the binary identifications of affect seen in figures 78-79. The incongruent combinations significantly lowered recognition and intensity rates of the high activation stimuli *happy* and *angry*, indicating a stronger role of the visual channel in the perception of these affects, but no such strong effect was seen for the low activation affects. Given the results of the congruent and the voice-only stimuli, we can say that the facial expression did have some effect, while still not being completely dominant over the voice.

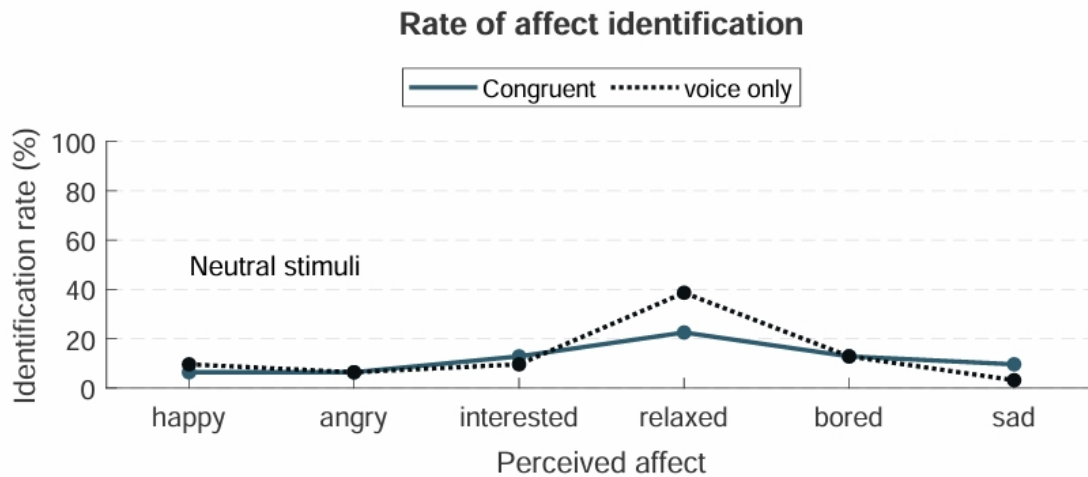


Figure 84 - Identification rates for the neutral stimuli in the visual context perception test.

To conclude, figure 84 illustrates the difference in performance for the neutral stimuli, presented in voice-only form and combined with the neutral visual stimulus. Here, the congruent visual had a very small effect in improving the perception of this vocal stimulus as a true neutral – the voice-only was more often perceived as *relaxed* than the congruent stimulus. Because the neutral was used as the reference level for the statistical analysis outlined above, the significance of this was not tested, but given the small difference seen in figure 84 it is unlikely to have been statistically significant.

7.2.3. Discussion and Conclusions

The initial hypothesis that the visuals would consistently strengthen the affect perception was not supported, as that effect was only seen for a couple of affects to a very small, not statistically significant degree. Still, this experiment also showed very high target affect identification rates, indicating that many of the voice-only stimuli were performing rather well and without much room for improvement for the identification of the congruent stimuli. The fact that the visual stimuli were intentionally kept very subtle may also have caused the strong dominance of the vocal affect seen in this experiment, especially for the *angry* voice, where the congruent stimuli instead lowered strength and target identification rates.

The addition of the visual context did seem to be slightly successful in aiding differentiation and unique signaling of a few affects compared to the voice-only stimuli,

especially for the *sad* and *relaxed* affects. While previous tests showed little to no differentiation between the *sad* and the *bored* stimuli, the addition of the congruent face, even as a subtle expression, appears to have improved the cueing of *sad* within the context of this experiment, even if not to a statistically significant degree. This, along with the visual validation experiment, aligns with evidence from the literature that suggests that some emotions are better recognized from the voice, and others from the face.

The lesser effect of the facial context can be seen in the *bored* affect, which, throughout this project, has been an elusive emotion to signal effectively. Once again, it is hypothesized that the unsuccessful performance of the *bored* affect is related to the absence of the creaky voice quality, as well as additional context cues that this affect may involve. The next experiment will further explore this differentiation between affects within both high and low activation categories, analyzing how dominant the face is when the visual context cues are swapped between affects of the same activation category.

The second part of the initial hypothesis, that the incongruent visuals would hinder recognition, was well demonstrated despite cases of clear dominance of the vocal affect; e.g., the *angry* voice seems to be particularly dominant. Still, the effect of the opposing facial expressions in lowering identification rates for the target vocal affects were very clearly shown for the high activation affects. These results may also indicate that low activation, negative valence affects are more strongly cued by the voice (overcoming any significant effects of incongruent combinations) and less malleable to the addition of context cues than the high activation affects seen here. However, it is important to note that this may also be a result of other factors, such as how the stimuli were created and analysis/testing methods. Nevertheless, each stimuli appears to mostly stay within its vocal target activation category, despite the opposing visual context.

With the interaction of the incongruent *relaxed* stimulus, this experiment also introduced the dimension of valence into the study, which was not previously considered in terms of affect differentiation. This may indicate that different information channels affect different dimensions of affect expression, which may be an interesting exploration for the future.

7.3. Experiment 4 – Visual Context with Within-Activation Congruence

The second experiment conducted with the combined vocal and visual stimuli is described in this subsection. The same logic regarding the choice for using only the combined vocal stimuli was applied. This experiment also aimed to investigate the effects of the visual stimuli on the affect recognition of the vocal stimuli, but combinations of incongruent stimuli were, this time, made within activation categories – i.e., low activation vocal stimuli (*sad*, *bored*, and *relaxed*) were combined incongruently with visuals for the other affects in the same activation category. The same applied to the high activation affects (*angry*, *happy*, and *interested*), with *interested* being only a visual stimulus combined with the *angry* and *happy* vocal stimuli. This was done based on results from the previous voice-only experiments, especially for the low activation affects, to explore if the face is an effective context cue to disambiguate those affects that were not uniquely cued within their activation categories.

The stimuli in this test also included the vocal stimuli on their own, and the congruent combinations between vocal and visual stimuli for comparison. A summary of the stimuli combinations can be found on table 31 below.

Table 31 - Vocal and visual stimuli combinations for the second perception test.

Vocal Stimulus	Visual Stimulus			
	None	Congruent	Incongruent 1	Incongruent 2
Angry	None	Angry	Happy	Interested
Bored	None	Bored	Relaxed	Sad
Happy	None	Happy	Angry	Interested
Relaxed	None	Relaxed	Bored	Sad
Sad	None	Sad	Bored	Relaxed
Neutral	None	Neutral	N/A	N/A

The total number of individual stimuli for this test was 22, including the two versions of neutral, presented in the same video format as the stimuli in subsection 7.2.

7.3.1. Perception Test Design and Subjects

This perception test was created in the same format as the one described in subsection 7.2 for comparability, but with a larger number of stimuli for listeners to judge. As previously, listeners were presented with instructions¹³, followed by primers of the vocal stimuli (the neutral individually, followed by all vocal affects). They were then presented with a practice run of 6 stimuli (which were discarded in the final analysis of results), followed by the proper three tasks described above, each video showing the neutral preceding the 22 stimuli presented in random order. This perception test was presented to 28 participants with varying levels of proficiency in Irish.

7.3.2. Statistical Analysis and Results

Statistical approaches and results of this test are presented below. For the statistical analysis, mixed-effects ordinal logistic regression models were used to measure the relationships between the intensity scores and target affects/visual stimulus combination types through R (R Core Team, 2021). The same model from experiment 3 was used but adapted to the new variables (additional incongruent conditions). Three separate models were fitted, one for each affect dimension present in the perception test: *sad-happy*, *bored-interested*, and *relaxed-angry*. The 7-point scale (between -3 and +3) used for participant ratings was treated as an ordinal dependent variable, with target affect and visual stimulus combination type being treated as independent variables, which were tested separately. The model also included the interactions between target affect and combination type, with random intercepts for listener ID to account for individual differences. Holm-Bonferroni adjustments (Holm, 1979) were also applied to account for multiple comparisons. Confidence intervals were calculated using standard errors from the model output, and *p* values below .05 after the Holm-Bonferroni adjustment were considered statistically significant. Reference categories used for the analysis were neutral for the target affect, and voice-only ('none') for the visual stimuli combination type. Model coefficients represent the log-odds of stimuli moving to a higher rating category on the ordinal scale (high activation affect categories), with negative values indicating a move towards the negative end of the scale (low activation categories).

¹³ Translations from English to Irish were done by Prof. Ailbhe Ní Chasaide.

Like in previous experiments, an overall model could not be fitted across all scales at once due to the nature of the experiment (i.e., only two of the target affects were available as perceived affects for each task), which caused numerical issues and quasi-complete separation of data during modelling. The per-scale model was therefore considered the most suitable and robust for the analysis presented. Since the experiment was conducted with a single measure to capture both affect identification and intensity rate, the model was written so that all significant values shown in the tables and figures indicate significance on both of those dimensions. Further clarification is given below. Results are presented in figures 85-90, and tables 33-35. As before, absent interactions were dropped due to rank deficiency that could cause coefficient estimation issues.

Figure 85 presents a heatmap confusion matrix showing listener's judgements of affect (perceived affect) per scale, based on each target affect and visual stimulus combination type, based only on a binary identification of affect (e.g., for the *sad-happy* scale, any positive values are counted as one instance of *happy* as the perceived affect, and negative values count as an instance of *sad*, and so on). Figure 86 presents an alternative representation of affect recognition rates for each affect-targeting quartet of combinations, also based on binary preprocessing. Figures 87-89 show the results of the ordinal regression analysis as confidence intervals over the coefficient estimates for the effects of target affect, visual stimulus combination type and their interactions for each test scale. Finally, figure 90 shows the confidence intervals for each combination quartet of affect-targeting stimuli on each scale, plotting the confidence intervals levelled to the original intensity scale of the perception test. Table 32 presents the definition of 'incongruent 1' and 'incongruent 2' for each affect, to be consulted alongside figure 90.

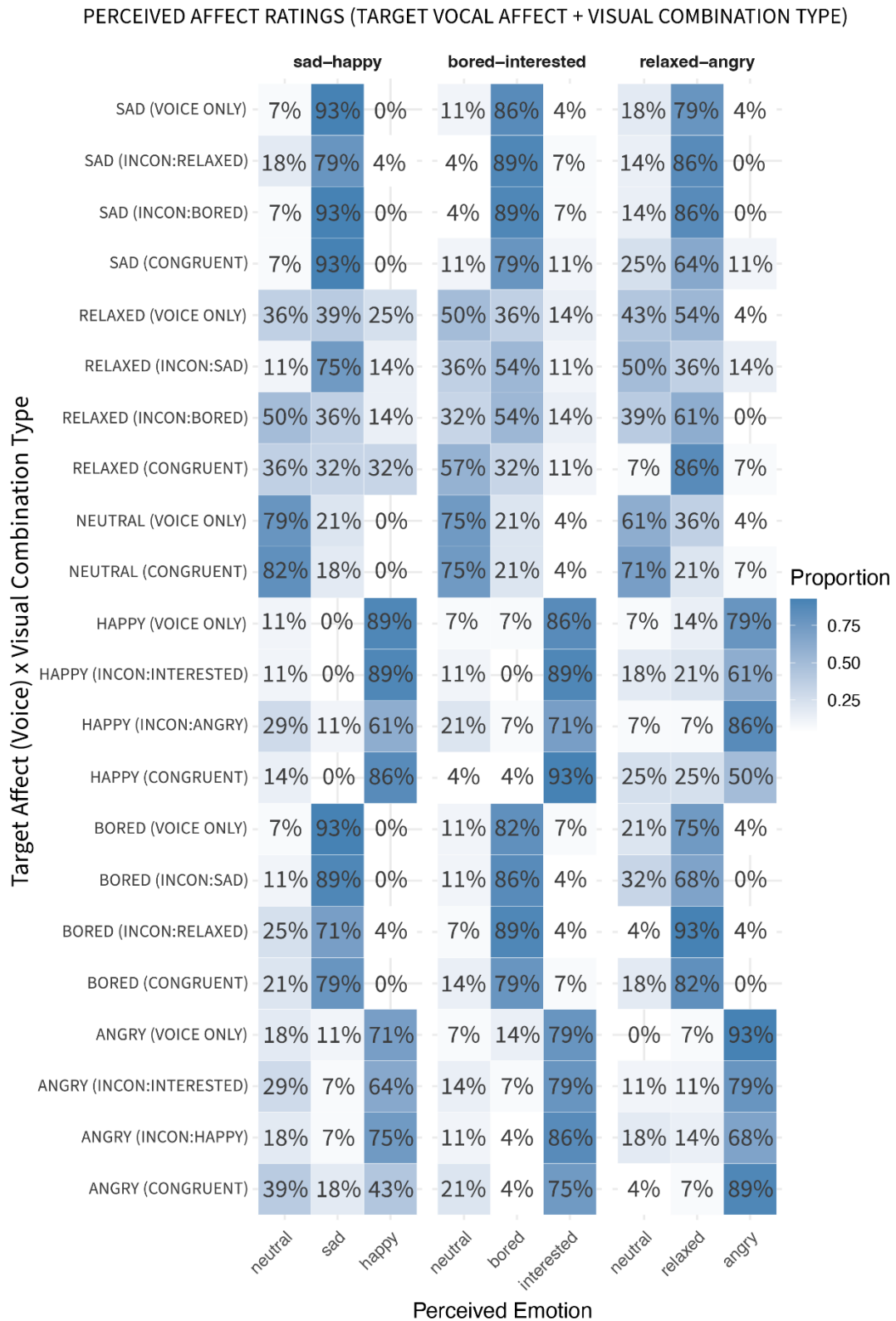


Figure 85 - Affect identification heatmap matrix showing the percentages of affect identification rates (perceived affect) based on target affect and visual combination types, and on binary identification of affect (i.e., considering only categorical identification on each end of the test scales, without the effects of intensity scores).

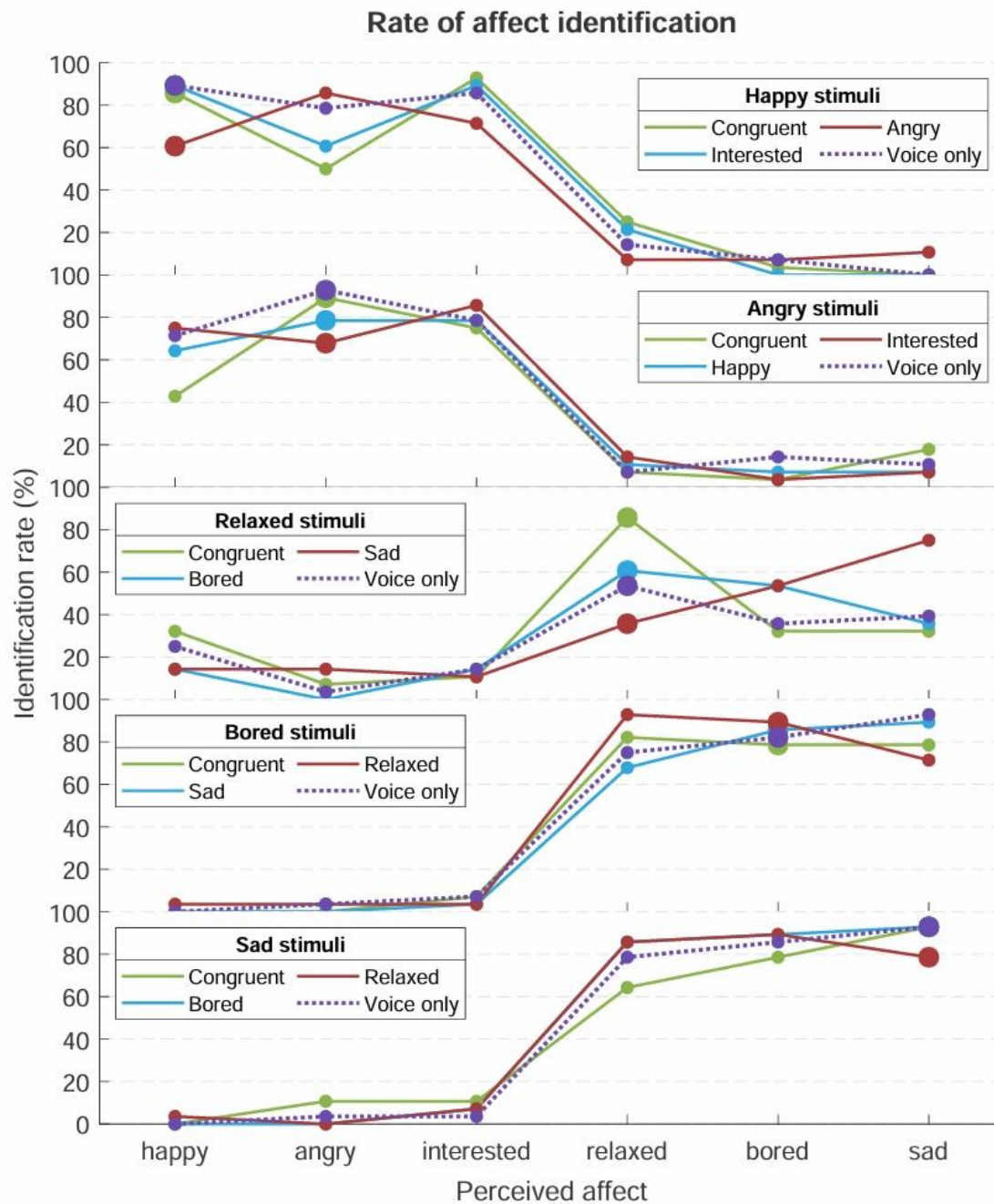


Figure 86 - Alternative representation of affect identification rates for each combination quartet of affect-targeting stimuli. Titles by each section of the image indicate the stimuli's target affect. Affect listings on the horizontal axis represent perceived affect from listeners, and the vertical axis represents percentages of identification rates. Identification rates are based on binary identification of affect (i.e., considering only categorical identification on each end of the test scales, without the effects of intensity scores). Neutral is shown as the dotted black line for comparison. The larger dots show the values of the target affect for each panel.

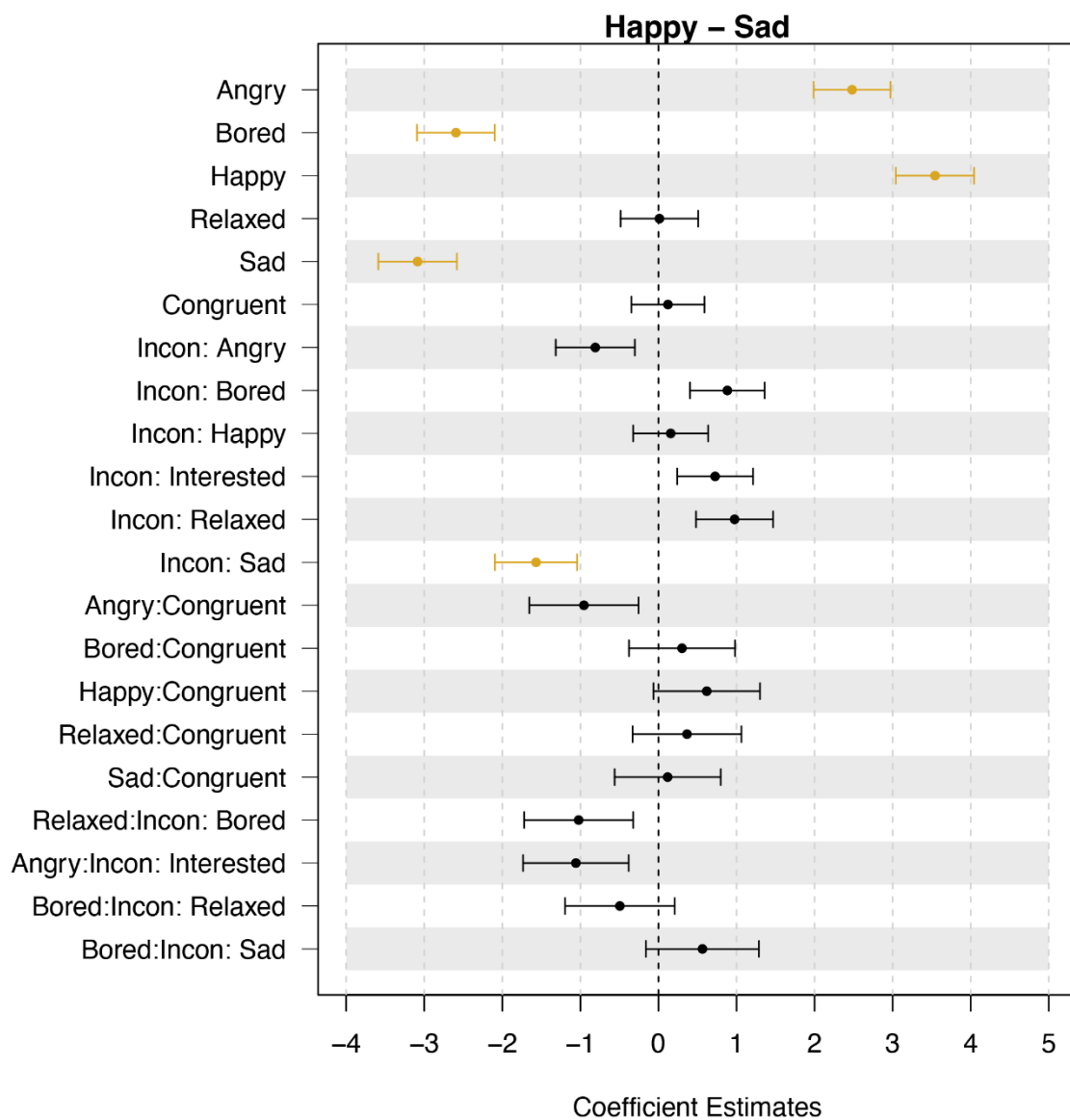


Figure 87 – Coefficient estimates for target affect, combination type and interaction affects from the mixed-effects ordinal regression analysis results for the *sad-happy* scale. Yellow lines indicate Holm-Bonferroni adjusted significances. Sidebars indicate Standard Error to each side. Bars completely above 0 on the horizontal axis indicates positive correlation, bars intersecting 0 indicate no correlation, and bars below 0 indicate negative correlation (see table 33 for numerical values).

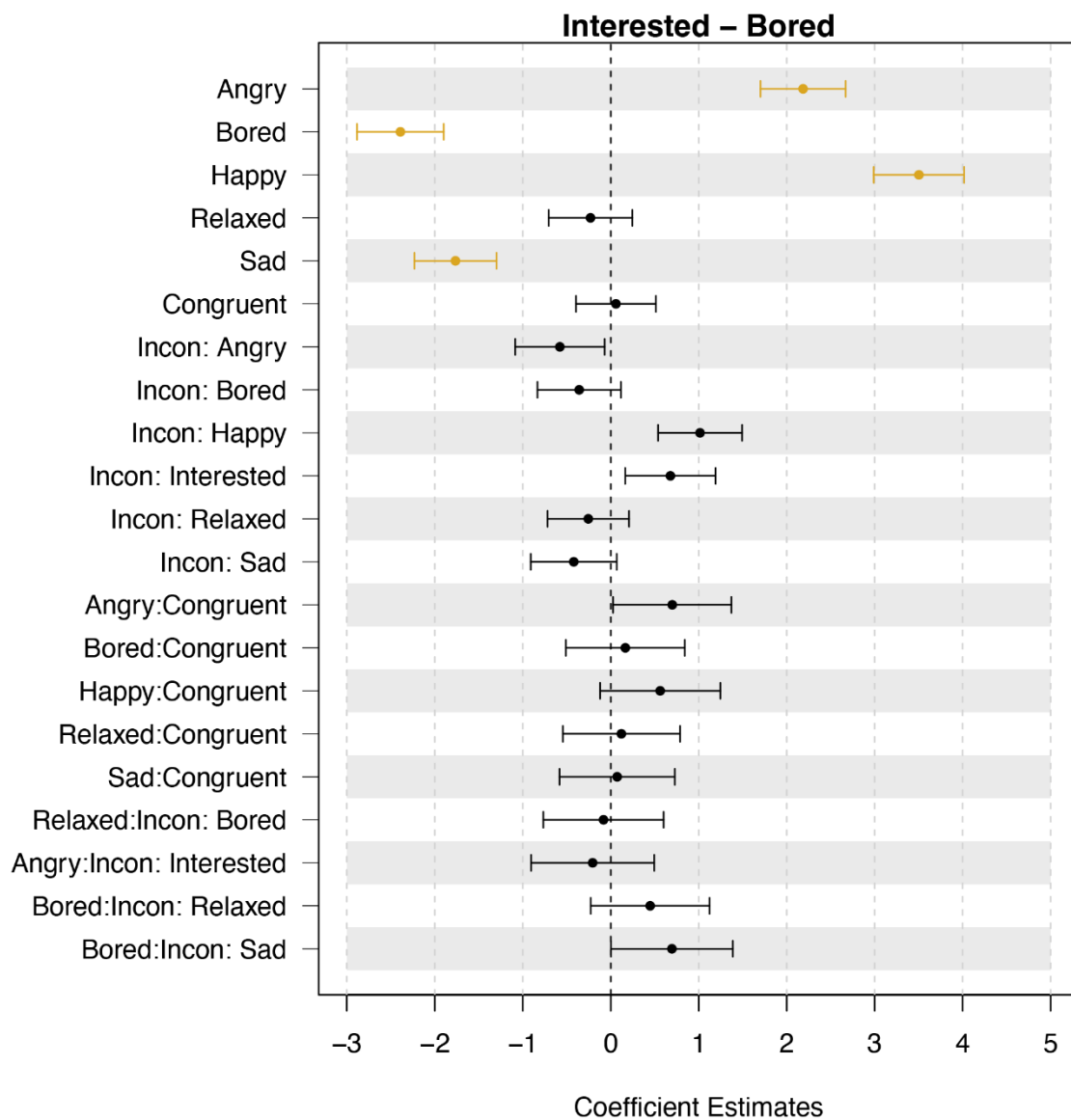


Figure 88 – Coefficient estimates for target affect, combination type and interaction affects from the mixed-effects ordinal regression analysis results for the *bored-interested* scale. Yellow lines indicate Holm-Bonferroni adjusted significances. Sidebars indicate Standard Error to each side. Bars completely above 0 on the horizontal axis indicates positive correlation, bars intersecting 0 indicate no correlation, and bars below 0 indicate negative correlation (see table 34 for numerical values).

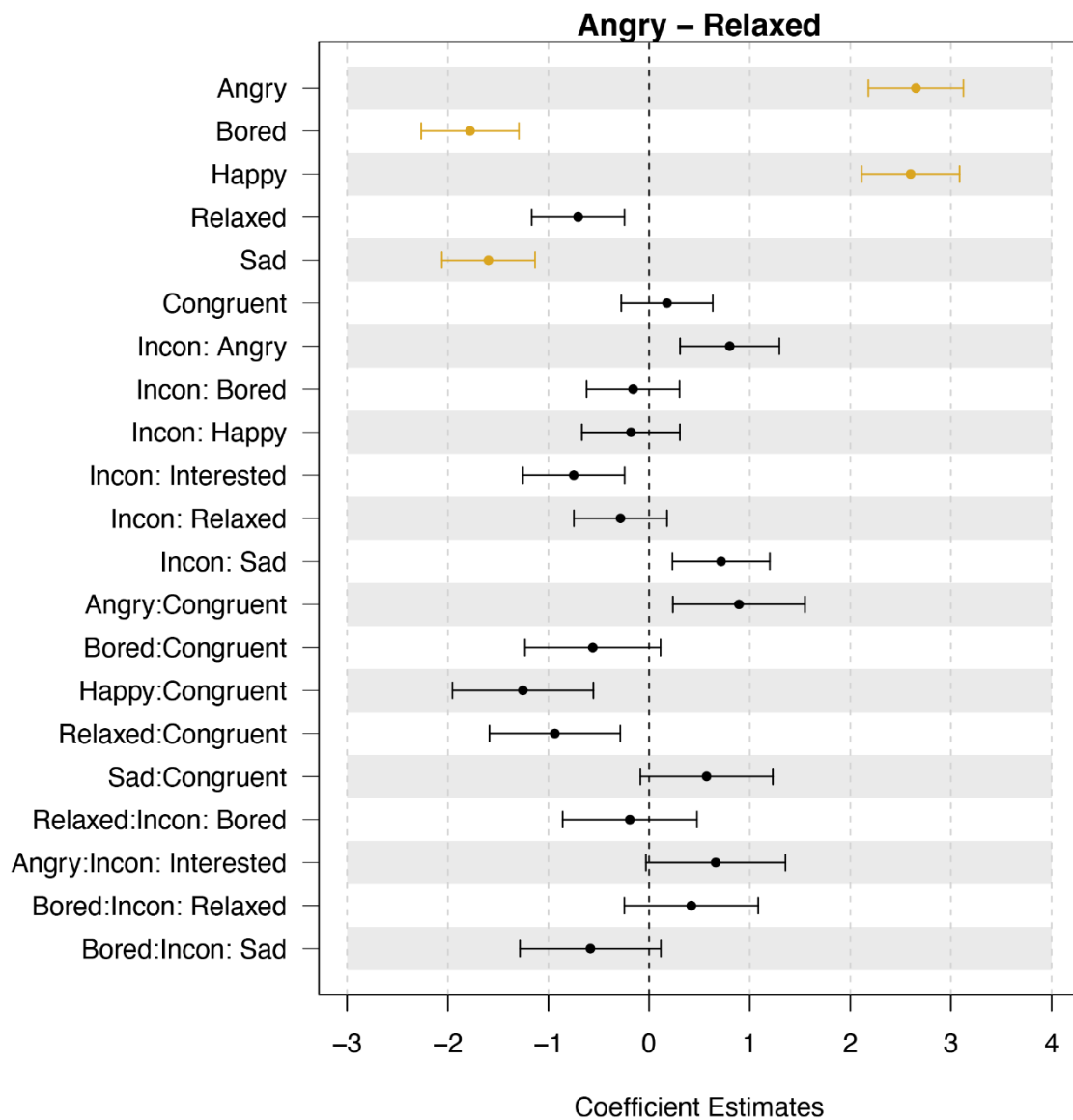


Figure 89 – Coefficient estimates for target affect, combination type and interaction effects from the mixed-effects ordinal regression analysis results for the *relaxed-angry* scale. Yellow lines indicate Holm-Bonferroni adjusted significances. Sidebars indicate Standard Error to each side. Bars completely above 0 on the horizontal axis indicates positive correlation, bars intersecting 0 indicate no correlation, and bars below 0 indicate negative correlation (see table 35 for numerical values).

Table 32 - Definitions of Incongruent 1 and Incongruent 2 for figure 90 analysis.

	<i>Incongruent 1</i>	<i>Incongruent 2</i>
<i>Angry</i>	Happy	Interested
<i>Happy</i>	Angry	Interested
<i>Bored</i>	Relaxed	Sad
<i>Sad</i>	Bored	Relaxed
<i>Relaxed</i>	Bored	Sad

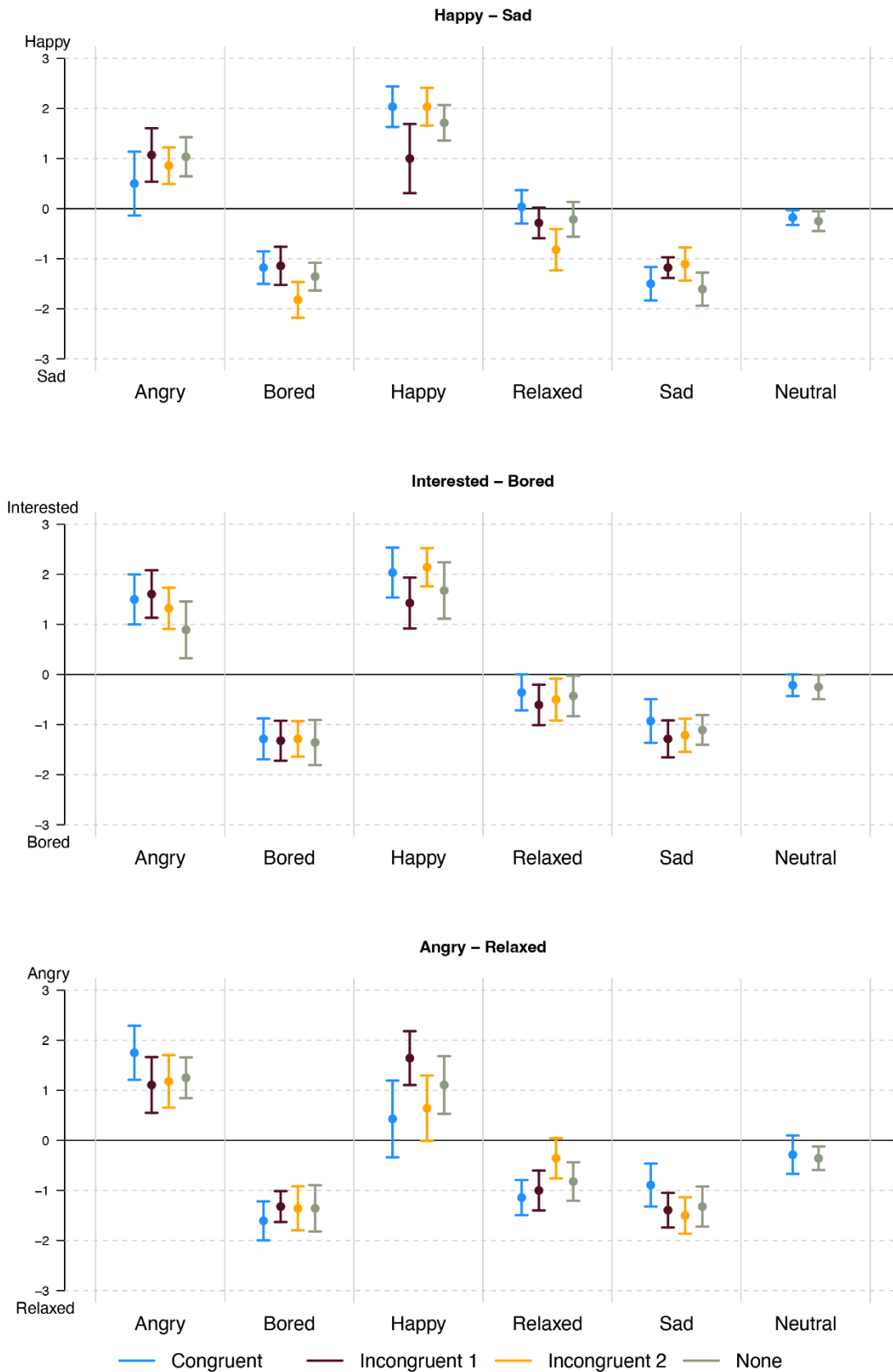


Figure 90 – Intensity score confidence Intervals for all target affects and quartets of visual combinations across all scales. Error bars indicate 95% confidence intervals (± 2 standard errors) around each mean.

Table 33 - Ordinal logistic regression results for the *sad-happy* scale with Holm-Bonferroni adjustments, showing main effects of target affect, visual stimulus combination type and interactions on listeners' perception of affect intensity. Positive estimates indicate positive correlation/above-chance performance relative to the reference categories; negative estimates indicate negative correlation/below-chance performance. Values close to 0 indicate no significant effect. Reference categories are *neutral* (target emotion) and *voice-only* (combination type). Absent interactions were dropped due to rank deficiency from showing patterns too similar to the reference levels. Significance levels: * $p < .05$, ** $p < .01$, *** $p < .001$.

**Target Emotion × Visual Stimulus Combination Type Effects on Intensity Ratings
(Sad-Happy Scale)**

<i>Predictor</i>	<i>Estimate (β)</i>	<i>SE</i>	<i>Z-value</i>	<i>P-value</i>	<i>P-value (Holm-Adjusted)</i>
Main Effects					
Target Affect: Angry	2.48	0.49	5.04	<.001***	<.001***
Target Affect: Happy	3.54	0.50	7.06	<.001***	<.001***
Target Affect: Bored	-2.59	0.50	-5.20	<.001***	<.001***
Target Affect: Sad	-3.09	0.50	-6.12	<.001***	<.001***
Target Affect: Relaxed	0.01	0.50	0.03	.98	1.00
Visuals: Congruent	0.12	0.47	0.26	.80	1.00
Visuals: Incon.Angry	-0.81	0.51	-1.60	.11	1.00
Visuals: Incon.Happy	0.16	0.48	0.33	.74	1.00
Visuals: Incon.Bored	0.88	0.48	1.85	.07	.97
Visuals: Incon. Interest	0.73	0.49	1.49	.14	1.00
Visuals: Incon. Sad	-1.57	0.53	-2.98	.003**	.05*
Visuals: Incon. Relaxed	0.98	0.49	1.98	.05*	0.77
Interactions					
Angry vs. Congruent	-0.95	0.70	-1.37	.17	1.00
Bored vs. Congruent	0.30	0.68	0.45	.66	1.00
Happy vs. Congruent	0.62	0.68	0.91	.36	1.00
Relaxed vs. Congruent	0.37	0.70	0.52	.60	1.00
Sad vs. Congruent	0.12	0.68	0.17	.86	1.00
Relaxed vs. Incon.Bored	-1.02	0.70	-1.46	.14	1.00
Angry vs. Incon. Interest	-1.06	0.68	-1.56	.12	1.00
Bored vs. Incon. Relaxed	-0.49	0.70	-0.70	.48	1.00
Bored vs. Incon.Sad	0.56	0.72	0.78	.44	1.00

Table 34 - Ordinal logistic regression results for the *bored-interested* scale with Holm-Bonferroni adjustments, showing main effects of target affect, visual stimulus combination type and interactions on listeners' perception of affect intensity. Positive estimates indicate positive correlation/above-chance performance relative to the reference categories; negative estimates indicate negative correlation/below-chance performance. Values close to 0 indicate no significant effect. Reference categories are *neutral* (target emotion) and *voice-only* (combination type). Absent interactions were dropped due to rank deficiency from showing patterns too similar to the reference levels.

Significance levels: * $p < .05$, ** $p < .01$, *** $p < .001$.

***Target Emotion × Visual Stimulus Combination Type Effects on Intensity Ratings
(Bored-Interested Scale)***

<i>Predictor</i>	<i>Estimate (β)</i>	<i>SE</i>	<i>Z-value</i>	<i>P-value</i>	<i>P-value (Holm-Adjusted)</i>
<i>Main Effects</i>					
Target Affect: Angry	2.19	0.48	4.52	<.001***	<.001***
Target Affect: Happy	3.50	0.52	6.81	<.001***	<.001***
Target Affect: Bored	-2.39	0.49	-4.85	<.001***	<.001***
Target Affect: Sad	-1.77	0.47	-3.79	<.001***	.003**
Target Affect: Relaxed	-0.23	0.48	-0.49	.63	1.00
Visuals: Congruent	0.06	0.46	0.13	.90	1.00
Visuals: Incon.Angry	-0.58	0.51	-1.14	.26	1.00
Visuals: Incon.Happy	1.02	0.48	2.13	.03*	0.57
Visuals: Incon.Bored	-0.36	0.47	-0.76	.45	1.00
Visuals: Incon. Interest	0.68	0.51	1.32	.19	1.00
Visuals: Incon. Sad	-0.42	0.49	-0.86	.39	1.00
Visuals: Incon. Relaxed	-0.26	0.46	-0.55	.58	1.00
<i>Interactions</i>					
Angry vs. Congruent	0.70	0.67	1.04	.30	1.00
Bored vs. Congruent	0.17	0.68	0.25	.82	1.00
Happy vs. Congruent	0.56	0.68	0.82	.41	1.00
Relaxed vs. Congruent	0.12	0.67	0.18	.86	1.00
Sad vs. Congruent	0.07	0.66	0.11	.91	1.00
Relaxed vs. Incon.Bored	-0.08	0.68	-0.12	.90	1.00
Angry vs. Incon. Interest	-0.21	0.70	-0.29	.77	1.00
Bored vs. Incon. Relaxed	0.45	0.68	0.66	.51	1.00
Bored vs. Incon. Sad	0.70	0.69	1.00	.32	1.00

Table 35 - Ordinal logistic regression results for the *relaxed-angry* scale with Holm-Bonferroni adjustments, showing main effects of target affect, visual stimulus combination type and interactions on listeners' perception of affect intensity. Positive estimates indicate positive correlation/above-chance performance relative to the reference categories; negative estimates indicate negative correlation/below-chance performance. Values close to 0 indicate no significant effect. Reference categories are *neutral* (target emotion) and *voice-only* (combination type). Absent interactions were dropped due to rank deficiency from showing patterns too similar to the reference levels. Significance levels: * $p < .05$, ** $p < .01$, *** $p < .001$.

**Target Emotion × Visual Stimulus Combination Type Effects on Intensity Ratings
(Relaxed-Angry Scale)**

Predictor	Estimate (β)	SE	Z-value	P-value	P-value (Holm-Adjusted)
Main Effects					
Target Affect: Angry	2.65	0.47	5.61	<.001***	<.001***
Target Affect: Happy	2.60	0.49	5.33	<.001***	<.001***
Target Affect: Bored	-1.78	0.49	-3.66	<.001***	.005**
Target Affect: Sad	-1.60	0.46	-3.44	.001**	.01*
Target Affect: Relaxed	-0.71	0.46	-1.53	.13	1.00
Visuals: Congruent	0.18	0.46	0.39	.70	1.00
Visuals: Incon.Angry	0.80	0.49	1.63	.10	1.00
Visuals: Incon.Happy	-0.18	0.49	-0.37	.71	1.00
Visuals: Incon.Bored	-0.16	0.46	-0.34	.73	1.00
Visuals: Incon. Interest	-0.75	0.51	-1.48	.14	1.00
Visuals: Incon. Sad	0.72	0.48	1.48	.14	1.00
Visuals: Incon. Relaxed	-0.28	0.46	-0.61	.54	1.00
Interactions					
Angry vs. Congruent	0.89	0.66	1.361	.17	1.00
Bored vs. Congruent	-0.56	0.67	-0.832	.41	1.00
Happy vs. Congruent	-1.25	0.70	-1.789	.07	1.00
Relaxed vs. Congruent	-0.94	0.65	-1.439	.15	1.00
Sad vs. Congruent	0.57	0.66	0.867	.39	1.00
Relaxed vs. Incon.Bored	-0.19	0.67	-0.286	.78	1.00
Angry vs. Incon. Interest	0.66	0.69	0.954	.34	1.00
Bored vs. Incon. Relaxed	0.42	0.67	0.631	.53	1.00
Bored vs. Incon. Sad	-0.58	0.70	-0.833	.41	1.00

Overall, the stimuli seem to be well defined in terms of activation categories, with no affects being perceived as its opposite activation category except for a few instances of the *relaxed* congruent stimulus being perceived as *happy* ($\beta = 0.01$, $p = .98$, figure 90). Results for the high activation affects seem very similar to the previous experiment for the voice-only and congruent stimuli, as would be expected.

The main effects of target affect and visual stimulus combination types can be seen in the upper halves of tables 33-35 and figures 87-89. For the target affects, strong and consistent effects were observed across all three scales. Participants consistently rated the stimuli higher on the intensity scales when the target affect aligned with the high activation affects (*happy* on the *sad-happy* scale, *angry* on the *relaxed-angry* scale, and both target affects as *interested* in the *bored-interested* scale), indicating a correlation between the high activation category and higher intensity ratings.

Specifically for the *sad-happy* scale, *happy*-targeting stimuli overall received the highest positive ratings ($\beta = 3.54$, $p < .001$), followed by *angry*-targeting stimuli ($\beta = 2.48$, $p < .001$) which were very often perceived as *happy* (see figures 85-87). For the low activation affects, the *sad*-targeting stimuli received the most negative ratings ($\beta = -3.09$, $p < .001$, illustrated in figure 90 in terms of smaller standard error bars), with *bored*-targeting stimuli being also rated negatively ($\beta = -2.59$, $p < .001$). *Relaxed*-targeting stimuli showed no significant difference from the neutral baseline.

On the *bored-interested* scale, the *happy*-targeting stimuli again showed the strongest positive effect ($\beta = 3.503$, $p < .001$), with the *angry*-targeting affects also rated positively ($\beta = 2.186$, $p < .001$). *Bored*-targeting stimuli showed the strongest negative effect ($\beta = -2.391$, $p < .001$), followed by the *sad*-targeting stimuli ($\beta = -1.765$, $p = .003$).

As for the *relaxed-angry* scale, *angry*-targeting stimuli received the highest ratings ($\beta = 2.652$, $p < .001$), followed closely by *happy*-targeting stimuli ($\beta = 2.599$, $p < .001$). Both the *bored* ($\beta = -1.779$, $p = .005$) and *sad*-targeting ($\beta = -1.596$, $p = .01$) stimuli were rated significantly lower than neutral, with results confirming their consistent ratings in the low activation end of all the scales.

In this experiment, the main effects of visual stimuli combination types were much more modest and less consistent across conditions than in the previous experiment, where

incongruent combinations were made across the two dimensions of the scale. Here, combining visual stimuli within the same activation category/across scales did not produce any significant effects compared to the voice-only reference level, and neither did the congruent combinations. This seems to suggest that the visual channel had very little effect on listener's ratings in both affect identification and intensity, and that the primary driver of listener judgements was the vocal channel.

One notable exception to this general rule was the addition of the incongruent *sad* visual which, as a main effect, did seem to sway the *bored* and *relaxed*-targeting stimuli further towards the more intensely *sad* end of the scale ($\beta = -1.57$, $p = .05$, table 33, figure 87). From figure 90, one may speculate that this effect came from adding the *sad* incongruent visual stimulus to the other low activation vocal affects on the *sad-happy* scale.

Although the ordinal regression model did test for interaction effects between the target affect categories and the visual stimuli combination type to see if the effect of the visual stimuli varied significantly for each target affect, none of these interactions reached statistical significance before or after the Holm-Bonferroni correction (all $p = 1.00$). This suggests that, this time, the effects of the visual stimuli combinations were consistent across different target affects, again with no significant difference found between the performance of the voice-only and the congruent stimuli in achieving better affect recognition and differentiation, and no large effects of the new types of incongruent combinations but for the small exception of the addition of the *sad* visual as an incongruent combination.

While not statistically significant, a few additional things to note emerged from the analysis and illustrations, mentioned here in order to inform potential changes to future experiments, given the results from the previous experiment (section 7.3) that indicate that the visual stimuli might have been too subtle in relation to the voice (even though the validation test showed high target identification rates).

In terms of the high activation affects, as mentioned, no difference was found in target identification rate or intensity scores between the congruent and the voice-only stimuli. The addition of the incongruent *interested* visual to the *happy* voice signalled the *happy* affect as much as the other two conditions. However, when added as an incongruent

visual to the *angry*-targeting voice, the *interested* visual slightly reduced its identification as *angry*, pulling it towards a more positive valence even more effectively than the addition of the incongruent *happy* visual stimulus to the *angry* voice (figure 86). On the *sad-happy* scale, the *angry* congruent stimulus was rated at a lower intensity than the other *angry* stimuli and the *happy* stimuli, which might indicate that the *angry* visual had a small effect on ratings (figures 85 and 86). Based only on the main effects, the *happy*-targeting stimuli was rated higher than the *angry*-targeting stimuli as *happy* on the *sad-happy* scale ($\beta = 3.54$ for *happy* and $\beta = 2.48$ for *angry*), and the *angry*-targeting stimuli were rated higher than the *happy*-targeting stimuli as *angry* than *happy* on the *relaxed-angry* scale ($\beta = 2.65$ for *angry*, $\beta = 2.60$ for *happy*), but the latter was a very small difference. If there is some small degree of differentiation, it would most likely come from what is seen in the *sad-happy* scale for the *angry* and *happy* congruent stimuli.

Additionally, in terms of binary identification, the voice-only *angry* appears to perform better than the congruent *angry* (figures 85-86), which we can again speculate to be an effect of the subtlety of the facial expression; however, the *relaxed-angry* scale in figure 90 shows the congruent *angry* stimulus as having a slight increase in intensity ratings, while also being pulled towards the negative valence side of the *sad-happy* scale. Other than these small effects mostly based on binary identification of the affective categories, however, there was no significant differentiation between *interested*, *happy* and *angry* affects for this experiment.

As before, the low activation affects showed no significant differences in target affect cueing and within activation category differentiation, and congruent combinations performed similarly, if not slightly worse, than the voice-only stimuli. The biggest difference in combination type performance was the incongruent *sad* visual stimulus applied to the *bored* and *relaxed*-targeting voices, as previously mentioned ($\beta = -1.57$, $p = 0.05$, figure 90). Overall, *sad*-targeting stimuli had a stronger effect than *bored*-targeting stimuli in their classifications on the *sad* end of the scale on the *sad-happy* scale ($\beta = -3.09$ for *sad*, $\beta = -2.59$ for *bored*), and *bored*-targeting stimuli had a stronger effect than the *sad*-targeting stimuli on the *bored-interested* scale ($\beta = -2.39$ for *bored*, $\beta = -1.77$ for *sad*), but there doesn't seem to be much variation from attaching the *bored* visual stimulus to the *sad* voice. Figure 90 doesn't indicate much difference in intensity ratings for these two affects either,

showing them both on the middle to lower end of the low activation sections. Therefore, the *bored* affect is, once again, elusive, and results are similar to those seen in the previous experiment. Overall, all the low activation affects seem to be almost equally signaled by the *bored*-targeting voice, and no combinations of the *bored* stimulus showed any signs of uniqueness. For the *sad*-targeting stimuli, the only stimulus that was slightly less frequently perceived as *sad* was the combination with the incongruent *relaxed* visual, with the voice still being dominant. No combinations of *sad*-targeting voice with any of the visual stimuli resulted in significant differences in the target affect identification rate.

The *relaxed*-targeting stimuli seem to mirror previous results with a slightly accentuated binary identification rate. This time, out of the three low activation stimuli, *relaxed* seems to be the most uniquely cued based on figure 86, in contrast to what was found in voice-only studies. The congruent stimulus seems to be particularly effective in improving the signaling this affect compared to its binary ratings for the other two affect options, but as mentioned, no significant differentiation was found from the neutral reference baseline.

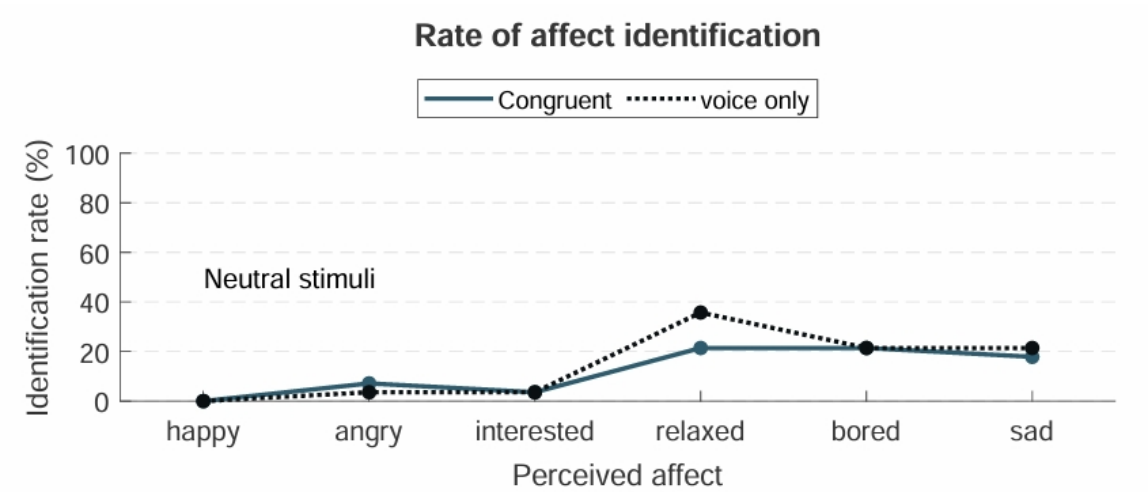


Figure 91 - Identification rates for the neutral stimulus in the second visual context perception test.

Finally, figure 91 shows the identification rate of the neutral stimuli for this test, presented only in voice-only and congruent form. Results are very similar to the previous experiment. The voice-only neutral, to some extent, cued the *relaxed* affect as well as *bored* and *sad*, while the congruent neutral stimulus reduced this ambiguity to a small degree.

Again, significance for this was not tested as the model used the neutral as the baseline for comparisons. The contour indicates that the neutral stimulus was very close to a true neutral.

7.3.3. Discussion and Conclusions

Similar conclusions to the previous experiment can be drawn here. While the effects of the congruent faces were not as strong as it was initially expected, this may be explained by the subtle nature of the facial expressions themselves. However, a small degree of differentiation appears to have emerged from the addition of the congruent visual stimuli through lowering the rates at which those stimuli were classified as other non-targeted affects within its activation category, which could indicate some potential for more intense versions of congruent facial expressions.

Compared to the previous experiment, the *sad* congruent stimulus showed slightly less affect-specific signaling here, while the *relaxed* did not have the same valence-pulling effect as previously seen in the *bored-interested* scale in experiment 3. The differences in the results for *sad* are harder to explain – it might be the case that the within-activation visual and vocal combinations caused some degree of fatigue on the listeners due to the perception test format (i.e., multiple stimuli being judged in sequence rather than spread out as in real life interactions), where the subtler low activation affects merged into each other; or it might simply be an effect of a slightly different pool of listeners, each having different personal contexts at the time the test was conducted.

The effects of the incongruent stimuli are varied, but all very subtle and non-significant with the exception of the *sad* visual added as an incongruent condition. The most obvious ones seen in figure 86 are the *happy*-targeting incongruent *angry* (*happy* voice with the incongruent *angry* face) and the *relaxed*-targeting incongruent *sad* (*relaxed* voice with the *sad* face), both of which showed the largest visual difference in binary identification rate from the congruent stimuli; and, to a smaller degree, the *angry*-targeting incongruent *interested* (*angry* voice with the *interested* face). Here, the face seems to have quite an effect on the perception of emotion in the voice, with the *interested* visual stimulus being relatively effective in signaling its respective affect. The successful cueing of *interested* by the high activation affects, despite there being no *interested*-targeting vocal stimuli,

continues to be a curious turn of events, both for the voice-only and the visual context studies.

While the congruent facial expressions did not perform much higher than the voice-only stimuli in terms of target identification rate, they once again seem to lessen the ambiguity within activation categories for certain affects seen in the voice-only studies. Given these results, it may be the case that facial expressions provide better differentiation of affects in terms of valence – such as differentiating *relaxed* from *bored* and *sad*, as well as *happy* from *angry* compared to the voice-only studies.

8. CONCLUSIONS

8.1. General Discussion and Conclusions

The aims of this thesis were, firstly, to explore the effectiveness of global and local manipulations of a small set of voice source parameters to evoke certain emotions. Secondly, using the same speech samples, it aimed to explore how the visual context impacts the perception of emotion.

The experiments were carried out in a very context-specific manner, by analyzing the vocal expression of emotion from a single speaker, within a single language and cultural context. The decision for the limited number of parameters was done with a future practical aim in mind – to apply the findings and methodology detailed here to Irish text-to-speech applications, hence the reasoning for maintaining relative simplicity in the parameter shifts.

The first contribution of this project was to develop a novel approach for simplifying the features of affective speech, stylizing the affective contours to achieve the desired simplicity but still effectively cue the target emotions. While the application of the global shifts drew on previous work by Haan et al. (1997), the methodology for applying local shifts, applied individually and on top of the global baseline, was developed specifically for the research hereby presented. The development of this approach is also closely linked to the development of the Voice Source Generator (the VSG, as described in Z. Wang & Gobl, 2023b), which was used in this project and extensively tested and refined in the course of this research. The results demonstrate that the stylization approach developed (and applied in the VSG system) was capable of modifying the neutral baseline utterance as desired, matching the contours of the production data, with good quality audio output. The test results indicate that the affect cueing was, to a degree, successful.

The first experiment, detailed in chapter 5, applied the analysis of affective production data and parameter manipulation methodology to a neutral baseline, resynthesized directly from the parametrization of a recorded neutral utterance, to produce a set of global, local, and combined stimuli for five discrete target affects (*happy*, *angry*, *bored*, *sad*, and *relaxed*). The hypothesis was that the combined stimuli would achieve the best target-affect recognition rates, followed by the local and global shifts in that order. The

results for this experiment were mixed. On the one hand, the local and combined stimuli performed the best for the *happy* affect, and *angry* was best signaled by the combined stimulus. It was, however, unclear how much the voice source transformations were responsible for the success of the *happy* signaling, given the additional tempo and formant shifts applied.

The low activation affects, on the other hand, were not as effectively cued and often classified as neutral, and there was no differentiation between the affects in this activation category. It was speculated that this was due to the fact that the intended neutral baseline recording appeared to be rather on the low activation side. Regardless, the primary hypothesis that the combined stimuli would be the most effective was not generally supported, and the secondary hypothesis (that the local stimuli would be more effective than the global stimuli) was only partly confirmed for the high activation affects.

The second experiment used the ABAIR synthesizer output to provide the neutral baseline. After presenting the stimuli to listeners in a different format than the previous test, the results showed the following.

For this experiment, the initial hypothesis was confirmed regarding the higher effectiveness of the combined stimuli. The global shifts also performed well, but only the *happy* affect was well differentiated within its activation category. The low activation affects remained ambivalent despite the shift in the neutral baseline – while their identification rate increased in terms of differentiation from the neutral affect, the low activation affects, overall, were not uniquely cued. In fact, *sad* was not uniquely cued from *bored* regardless of the additional tempo shift applied to this affect, adding to the part of the literature that suggests that boredom can also be cued through slower tempo.

As mentioned, the surprising part was that the local stimuli performed much worse than expected based on previous experiments. This was possibly due to the shift of the neutral baseline, or indicates the necessity of testing more dramatic local shifts in future experiments. Regardless, this experiment clearly indicated that local shifts applied on their own to a neutral baseline were not as effective as the global shifts in the cueing of affect, but both target affect and intensity ratings were improved when both global and local shifts were combined.

While the change in the neutral baseline and perception test methodology was changed between the two tests and therefore complicates a direct comparison between the two, a few conclusions can be drawn from these voice-only stimuli. Results show that the low activation affects are not clearly differentiated by the stimuli presented, whereas the high activation affects *happy* and *angry* were both more successfully cued and *happy* showed more differentiated signaling.

Studies and reviews on vocal and facial expression of emotion also speculate that some emotions might be better identified from cues other than the voice, i.e., different channels are more crucial for recognition of different emotions than others (e.g., Johnstone & Scherer, 2000; Juslin et al., 2021; Keltner et al., 2016; Kraus, 2017; Pell, 2002; Scherer, 2003). Following this line of thought, the second set of experiments involving the visual cues was conducted based on the hypothesis that additional context can not only improve recognition of emotion from the voice, but also that the face may be particularly effective in differentiating emotions that have relatively similar acoustic profiles (e.g., among low or high activation affects).

The visual context experiments show that the addition of the visual stimuli did not have as big of a positive effect as expected in terms of affect recognition rate. It did, however, slightly improve more differentiated emotion signaling in terms of valence within the affect's activation category. This difference is seen particularly for the *sad* affect in the first experiment of the set, for *relaxed* in certain scales, and for the high activation affects *angry* and *happy*.

The incongruent visual stimuli showed mixed effects. When combinations were made across activation categories or across valence categories, the perception of the affect was drawn away from its target vocal affect. However, the effects of vocal target affects were still stronger than the effects of the visual stimuli. *Relaxed* was the one affect where the incongruent opposite stimulus (*angry* visual) almost caused ambiguity in its recognition.

When the incongruent facial expressions were of the same activation category, they mostly seemed to lead to differentiation in terms of valence – emotions of the same valence were often ambiguous, while those of opposite valence were slightly more well separated.

The exception was the case of *sad* in the second experiment, where the *sad* facial expression was particularly effective in the *sad-happy* scale as dominant over the other visual stimuli.

Signaling of the *interested* affect was often associated with the *angry* and *happy*-targeting stimuli. Because the visual stimulus was of positive valence, the addition of the incongruent *interested* visual to the *angry* affect did lower its identification rate as *angry*, and the *happy*-targeting stimuli were more often identified as *interested* than the *angry* stimuli; however, the rate of the latter was still significantly high. This may be because, when speakers are interested, it can be an emotion of positive or negative valence (i.e., one can be interested in something because it makes them excited and happy, or because it causes them to be angry). As previously mentioned, further experimentation would be needed to see if the cueing of *interested* could be achieved without the undertones of *angry* and *happy*, and it may be characterized by large local shifts in the contour.

The use of subtle facial expressions might have lessened the effect of the visual context, especially for the *angry*-targeting combinations. Results might have changed if the expressions were stronger, both in terms of identification rate as well as differentiation between affects. Still, while the high activation affects were much more successfully and uniquely identified from the voice than the low activation ones, the facial validation experiment conducted in chapter 7 showed that, for the low activation affects, the face was an important and effective cue for emotion identification – particularly for the *sad* affect. This may suggest that not only are additional cues important in the differentiation of emotion, but also that each information channel might aid with differentiation of affects through different dimensions, such as valence for the face and activation for the voice. Studies utilizing dimensional approaches could therefore shed light into the effectiveness of each channel into the perception of emotion in the pursuit of finding more effective, unique profiles for emotion categories or gradients.

It can also be speculated that, as mentioned, each affect might rely more heavily on a different channel in the same way that each affect might be more effectively cued through different kinds of changes to vocal parameters (global for low activation affects, local for high activation, and the potential of creaky voice for cueing *bored*). For example, *relaxed* benefitted from the addition of the congruent face in contrast to the voice-only stimulus in the visual context experiment. This time, only voices and faces were explored, but gestures

and body language, along with further changes to the voice, might be key to further unique signaling of other affects. Additionally, it must be said that general context, such as situation and familiarity with speaker's baseline and personality, may also be a big factor in accurate emotion identification as suggested by the literature.

Finally, it appears that the combined stimuli are relatively successful at achieving target affect identification across most affects (*bored* being the one most in need of further investigation) if the system of modifications proposed here is to be used in the suggested applications. Further refining and experimentation during implementation would be needed along with evaluation of its effectiveness in the full context of the applications' use, but it is encouraging to see that some affects were particularly well signaled – especially the very elusive *happy*.

8.2. Limitations of the Study

The limitations of the present study are detailed here. First, it should be said that this research does not set out to fully explore all voice source parameters related to affect, as the number of parameters manipulated was intentionally very limited. For example, for the *bored* affect, creaky voice, which is often found to be correlated to boredom in the literature, was not included here. There is scope for further investigation of other voice source parameters as well as spectral characteristics for a more complete picture of how emotion is expressed through the voice. The methodology presented here only aimed to be able to achieve the signaling of a selected number of affect categories very commonly used in day-to-day interactions, with a view to being able to manipulate the existing synthetic voices to allow users to add some degree of affective coloring to the message.

The findings in this study also only apply to declarative sentences. As prosodic patterns are known to be different for questions and different types of answers, it would be interesting for future follow-up studies to explore the effects of emotion on other sentence types.

Because the production data were acquired through acted portrayals, the analyzed utterances are also not necessarily representative of nuanced, real every-day emotion, being more stereotypical portrayals to be extrapolated to speech synthesis, and with the

desired practical implementations in mind. Furthermore, the production data were also acquired from a single speaker – the same speaker featured as one of the ABAIR voices. Clearly, there is no single way to express a specific emotion – but the goal here was to capture, insofar as possible, one possible way in which an affect is expressed with a view to enable some level of affect expression in the synthetic voice in the future. It would be of interest to explore the effectiveness of these voice manipulations on other Irish ABAIR voices, particularly female voices.

Additionally, the decision for using manual voice source analysis methods also meant that this analysis was only conducted for a handful of stimuli from a single utterance for most affects. This was done for reasons previously stated. Automatic voice source estimation methods available that also provided source parametrization through the LF model do not perform at a satisfying level at the time of writing, requiring a trade-off between the potentially inaccurate analysis of a larger set of data, or a smaller set of data analyzed to a gold standard. The decision for quality over quantity was a conscious one, but must still be mentioned.

Another issue identified later in the research was the lack of pre-validation of the production data before conducting voice source analysis. Ideally, the recorded stimuli would have been tested against a pool of listeners to evaluate if the acted portrayals were representative of the target affects by the listeners, rather than solely trusting the judgement of the actor and conducting informal listening sessions after the fact.

The perception test format also comes with its own issues. As mentioned by Bryant (2021), this kind of empirical approach is not reflective of a natural, social context, and “making repeated judgements of vocal clips out of (a wider) context only triggers the appropriate detection systems to a limited extent” (p.30), i.e., the listener is intentionally following instructions to identify emotion in isolated audio stimuli rather than that judgement being naturally triggered. Additional variables such as the entirety of the research context, the emotional and personal state of the listener while engaging in the research task, and the format of the task itself can also influence the results to a certain extent. There are, however, few options as to how to evaluate emotion in an empirical setting, and thus such testing limitations are a feature of any study in the field.

As covered in the review of the literature, there is also the topic of emotion models for vocal affect research. Here, the discrete model of emotion was used for a small number of emotion categories as it best fit the needs of the study and available resources. However, this approach is not exempt of criticism, and therefore it must be mentioned that some researchers prefer the dimensional models of emotion.

While the visual stimuli were made as realistic as possible to avoid distractions, they were created within certain time and software constraints. The animations were rendered to the best possible quality and realism and provided a more natural setting for the vocal stimuli to be tested than static images would, but were still created in a very systematic manner and are still not completely representative of real-life interactions.

8.3. Future Directions

There are many aspects of this thesis that could benefit from further, more focused research. First is the need for faster and more accurate analysis tools that can handle a larger amount of recorded data to provide a richer set of production parameters for the exploration of affect in a wider variety of voices and Irish dialects. It would be particularly beneficial to reproduce the studies presented here using a female voice, or a different Irish dialect to see how shifts related to emotion may vary between those and the speaker in this thesis. The Voice Source Generator is currently being developed to allow for automatic inverse filtering, which seems promising given its capabilities seen in this project.

Automatic voice source analysis of more stimuli and a wider variety of sentences would also be critical for testing the shifts proposed in this thesis' methodology, and potentially refine the system used here by adding more detail to each affect. Additionally, more attention could be given to other voice parameters or features of voice quality, should systems be modified to allow for their manipulation. It would be interesting to, in the future, add creaky voice to the *bored* stimuli presented here to see how it affects recognition in the synthetic voice.

More attention may also be given individually to the addition of context cues to the emotional voice stimuli. The visual stimuli presented here were subtle and had a small effect on the identification rate of the vocal affects, but were still relatively successful in

differentiating those affects within their activation categories. Future research could potentially explore different and more intense versions of these visual stimuli. Likewise, should the analysis and manipulation of sentences be made faster and more convenient, the semantic context could also be explored by creating semantically emotional sentences for each affect and switched between affects, as done with the visual stimuli.

Finally, it would be interesting to see how the shifts proposed here would perform when systematically implemented in the ABAIR TTS synthesizer and the applications that currently use the ABAIR voices, once more developments arise. Of special interest for the application of findings is the Geabaire system (Barnes et al., 2023), in order to allow users to add expressiveness to their synthesized input when desired.

REFERENCES

- Abe, M., Nakamura, S., Shikano, K., & Kuwabara, H. (1990). Voice conversion through vector quantization. *Journal of the Acoustical Society of Japan (E)*, 11(2), 71-76. doi:10.1250/ast.11.71
- Abelin, Å. (2004). *Spanish and Swedish interpretations of Spanish and Swedish emotions - the influence of facial expressions*. Paper presented at 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 ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion, Newcastle, Northern Ireland
- Adobe Inc. (2023). Adobe Photoshop (Version 24.7.5): Adobe Inc. [Computer Software]. <https://www.adobe.com/products/photoshop.html>
- Aicha, A. B. (2018). Voice source modelling using modified LF model with reduced parameters. *Procedia Computer Science*, 126, 423-430. doi:<https://doi.org/10.1016/j.procs.2018.07.276>
- Aihara, R., Takashima, R., Takiguchi, T., & Ariki, Y. (2012). GMM-Based Emotional Voice Conversion Using Spectrum and Prosody Features. *American Journal of Signal Processing*, 2(5), 134-138.
- 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 parametrisation*. Paper presented at INTERSPEECH 2005, Lisbon, Portugal.
- Akuzawa, K., Iwasawa, Y., & Matsuo, Y. (2018). *Expressive Speech Synthesis via Modeling Expressions with Variational Autoencoder*. Paper presented at INTERSPEECH 2018, Hyderabad, India.
- 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 parametrization of the glottal flow. *The Journal of the Acoustical Society of America*, 112(2), 701-710.
- Alku, P., Pohjalainen, J., Vainio, M., Laukkanen, A.-M., & Story, B. H. (2013). Formant frequency estimation of high-pitched vowels using weighted linear prediction. *The Journal of the Acoustical Society of America*, 134(2), 1295-1313. doi:10.1121/1.4812756
- Alku, P., & Vilkman, E. (1994). *Estimation of the glottal pulseform based on discrete all-pole modeling*. Paper presented at the 3rd International Conference on Spoken Language Processing (ICSLP 1994), Yokohama, Japan.

- 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.
- Alpert, M., Kurtzberg, R. L., & Friendhoff, A. J. (1963). Transient voice changes associated with emotional stimuli. *Archives of General Psychiatry*, 8(4), 362-365.
doi:<https://doi.org/10.1001/archpsyc.1963.01720100052006>
- Altenmüller, E., Schmidt, S., & Zimmermann, E. (2013). *The evolution of emotional communication: From sounds in nonhuman mammals to speech and music in man*. Oxford, UK: Oxford University Press.
- Andersen, P. A., & Guerrero, L. K. (1998). *Handbook of communication and emotion: Research, Theory, Applications, and Contexts*. New York: Academic Press.
- Arias, P., Belin, P., & Aucouturier, J.-J. (2018). Auditory smiles trigger unconscious facial imitation. *Current Biology*, 28(14), R782-R783. doi:[10.1016/j.cub.2018.05.084](https://doi.org/10.1016/j.cub.2018.05.084)
- Arias, P., Rachman, L., Liuni, M., & Aucouturier, J.-J. (2021). Beyond Correlation: Acoustic Transformation Methods for the Experimental Study of Emotional Voice and Speech. *Emotion Review*, 13(1), 12-24. doi:<https://doi.org/10.1177/1754073920934544>
- Arias, P., Soladie, C., Bouafif, O., Roebel, A., Segulier, R., & Aucouturier, J.-J. (2020). Realistic Transformation of Facial and Vocal Smiles in Real-Time Audiovisual Streams. *IEEE Transactions on Affective Computing*, PP(99), 1-1. doi:[10.1109/taffc.2018.2811465](https://doi.org/10.1109/taffc.2018.2811465)
- Arnold, M. B. (1960). *Emotion and personality* (Vol. 1 & 2). New York: Columbia University Press.
- Arvaniti, A. (2020). The Phonetics of Prosody. In *Oxford Research Encyclopedia of Linguistics*: Oxford University Press.
- Aubergé, V., & Cathiard, M. (2003). Can we hear the prosody of smile? *Speech Communication*, 40(1), 87-97.
- Audacity Team. (2021). Audacity(R): Free Audio Editor and Recorder (Version 3.0.0) [Computer Software]. <https://audacityteam.org/>
- 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).
- 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. ACII 2005. Lecture Notes in Computer Science* (Vol. 3874, pp. 527-534). Berlin, Heidelberg: Springer.
- Audibert, N., Aubergé, V., & Rilliard, A. (2010). Prosodic correlates of acted vs. spontaneous discrimination of expressive speech: A pilot study. *Speech Prosody 2010*, 1000979, 1-4.
- Bachorowski, J.-A. (1999). Vocal expression and perception of emotion. *Current Directions in Psychological Science*, 8(2), 53-57.

- Bachorowski, J.-A., & Owren, M. J. (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. In *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* (Vol. 244-265).
- Bachorowski, J.-A., & Owren, M. J. (2008). Emotion-related vocal acoustics. Cue configuration, dimensional, and affect-induction perspectives. In K. Izbebski (Ed.), *Emotions in the Human Voice* (Vol. 1, pp. 87-99). San Diego, CA: Plural Publishing.
- Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70(3), 614-636.
- Bänziger, T., Hosoya, G., & Scherer, K. R. (2015). Path Models of Vocal Emotion Communication. *PLoS ONE*, 10(9), e0136675. doi:10.1371/journal.pone.0136675
- 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 8th European Conference on Speech Communication and Technology, 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 expressions. *Speech Communication*, 46(3), 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.
- Barakat, H., Turk, O., & Demiroglu, C. (2024). Deep learning-based expressive speech synthesis: a systematic review of approaches, challenges, and resources. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024(11). doi:https://doi.org/10.1186/s13636-024-00329-7
- Barkhuysen, P., Krahmer, E., & Swerts, M. (2007). *Cross-modal perception of emotional speech.* Paper presented at the 16th International Congress of Phonetic Sciences (ICPhS XVI), Saarbrücken, Germany.
- Barnes, E., Cummins, J., Errity, R., Morrin, O., Berthelsen, H., Wendler, C., . . . Ní Chasaide, A. (2023). *Geabaire, the First Irish AAC System: Voice as a Vehicle for Change* Paper presented at the 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023), Dublin, Ireland.
- Barnes, E., Morrin, O., Ní Chasaide, A., Cummins, J., Berthelsen, H., Murphy, A., . . . Ní Chiaráin, N. (2022). *ACC don Ghaeilge: the Prototype Development of Speech-Generating Assistive Technology for Irish.* Paper presented at the 4th Celtic Language Technology Workshop within LREC2022, Marseille, France.
- Barrett, J., & Paus, T. (2002). Affect-induced changes in speech production. *Experimental Brain Research*, 146(4), 531-537.

- Barrett, L. F. (1998). Discrete emotions or dimensions? The role of valence focus and arousal focus. *Cognition & Emotion*, 12, 579-599.
- Barrett, L. F., Mesquita, B., & Gendron, M. (2011). Context in Emotion Perception. *Current Directions in Psychological Science*, 20(5), 286-290. doi:10.1177/0963721411422522
- Bates, D. M., Martin, B., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48. doi:10.18637/jss.v067.i01
- 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., 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, Marrakesh, Morocco.
- Beaucousin, V., Lacheret, A., Turbelin, M.-R., Morel, M., Mazoyer, B., & Tzourio-Mazoyer, N. (2007). fMRI Study of Emotional Speech Comprehension. *Cerebral Cortex*, 17(2), 339-352. doi:10.1093/cercor/bhj151
- Beckman, M. E., & Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology*, 3, 255-309. doi:10.1017/S095267570000066X
- Beier, E. G., & Zautra, A. J. (1972). Identification of vocal communication of emotion across cultures. *Journal of Consulting and Clinical Psychology*, 39, 166.
- Ben-David, B. M., Multani, N., Shakuf, V., Rudzicz, F., & Lieshout, P. H. H. M. v. (2016). Prosody and Semantics Are Separate but Not Separable Channels in the Perception of Emotional Speech: Test for Rating of Emotions in Speech. *Journal of Speech, Language, and Hearing Research*, 59(1), 72-89. doi:10.1044/2015_JSLHR-H-14-0323
- Bird, E., & Garellek, M. (2019). *Dynamics of voice quality over the course of the English utterance*. Paper presented at the Proceedings of the 19th International Congress of Phonetic Sciences.
- Birkholz, P. (2013). Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PLoS ONE*, 8(4), e60603. doi:10.1371/journal.pone.0060603
- Birkholz, P., Martin, L., Willmes, K., Kröger, B. J., & Neuschaefer-Rube, C. (2015). The contribution of phonation type to the perception of vocal emotions in German: An articulatory synthesis study. *The Journal of the Acoustical Society of America*, 137(3), 1503-1512. doi:10.1121/1.4906836
- Boersma, P., & Weenink, D. (2002). Praat, a system for doing phonetics by computer. *Glott International*, 5(9-10), 341-345.
- Boersma, P., & Weenink, D. (2024). Praat: doing phonetics by computer (Version 6.4.05) [Computer Software]. <http://www.praat.org/>
- Bóhm, T., Audibert, N., Shattuck-Hufnagel, S., Németh, G., & Aubergé, V. (2008). Transforming modal voice into irregular voice by amplitude scaling of individual glottal cycles. *The Journal of the Acoustical Society of America*, 123(5), 6141-6146. doi:10.1121/1.2935816

- Bonada, J., & Blaauw, M. (2013). *Generation of growl-type voice qualities by spectral morphing*. Paper presented at the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, New York, NY.
- Borod, J. C., Pick, L. H., Hall, S., Sliwinski, M., Madigan, N., Obler, L. K., . . . Tabert, M. (2000). Relationships among Facial, Prosodic, and Lexical Channels of Emotional Perceptual Processing. *Cognition and Emotion*, 14(2), 193-211. doi:10.1080/026999300378932
- 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 & Emotion*, 15(1), 57-79. doi:10.1080/0269993004200114
- Broad, D. J. (1979). The New Theories of Vocal Fold Vibration. In N. J. Lass (Ed.), *Speech and Language* (Vol. 2, pp. 203-256): Elsevier.
- Brunswik. (1956). *Perception and the representative design of psychological experiments*. Berkeley, CA: University of California Press.
- Bryant, G. A. (2021). The Evolution of Human Vocal Emotion. *Emotion Review*, 13(No. 1), 25-33. doi:https://doi.org/10.1177/1754073920930791
- Bryant, G. A., & Aktipis, C. A. (2014). The animal nature of spontaneous human laughter. *Evolution and Human Behavior*, 35(4), 327-335. doi:https://doi.org/10.1016/j.evolhumbehav.2014.03.003.
- Bryant, G. A., & Barrett, H. C. (2008). Vocal Emotion Recognition Across Disparate Cultures. *Journal of Cognition and Culture*, 8(1-2), 135-148. doi:https://doi.org/10.1163/156770908X289242
- Buck, R. (1984). *The communication of emotion*. New York: Guilford Press.
- Burkhardt, F., Audibert, N., Malatesta, L., Türk, O., Arslan, L. M., & Aubergé, V. (2006). *Emotional prosody - does culture make a difference?* Paper presented at Speech Prosody 2006, Dresden, Germany.
- Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). *A database of German emotional speech*. Paper presented at INTERSPEECH 2005, Lisbon, Portugal.
- Burkhardt, F., & Sendlmeier, W. F. (2000). *Verification of acoustical correlates of emotional speech using formant-synthesis*. Paper presented at the ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion, Newcastle, Northern Ireland.
- Burred, J. J., Ponsot, E., Goupil, L., Liuni, M., & Aucouturier, J.-J. (2019). CLEESE: An open-source audio-transformation toolbox for data-driven experiments in speech and music cognition. *PLoS ONE*, 14(4), e0205943. doi:10.1371/journal.pone.0205943
- Buss, D. M. (1995). Evolutionary psychology: A new paradigm for psychological science. *Psychological Inquiry*, 6, 1-30.
- Buss, D. M. (2020). Evolutionary Psychology Is a Scientific Revolution. *Evolutionary Behavioral Sciences*, 14(4), 316-323.

- Cahn, J. (1990). Generation of Affect in Synthesized Speech. *Journal of the American Voice I/O Society*, 8, 1-19.
- Callier, P. (2010). *Voice quality, rhythm and valorized femininities*. Paper presented at the Sociolinguistics Symposium 18, Southampton, UK.
- Callier, P. (2013). *Linguistic Context and the Social Meaning of Voice Quality Variation*. (PhD Thesis). Georgetown University.
- Campbell, N. (2005). Developments in corpus-based speech synthesis: approaching natural conversational speech. *IEICE Transactions on Information and Systems*, E88-D(3), 376-383. doi:10.1093/ietisy/e88-d.3.376
- Campbell, N., & Mokhtari, P. (2003). *Voice quality: the 4th prosodic dimension*. Paper presented at the 15th International Congress of Phonetic Sciences (ICPhS XV) Barcelona, Spain.
- Cannon, W. (1929). *Bodily changes in pain, hunger, fear and rage*. New York: Appleton.
- Caridakis, G., Malatesta, L., Kessous, L., Amir, N., Raouzaoui, A., & Karpouzis, K. (2006). *Modelling naturalistic affective states via facial and vocal expression recognition*. Paper presented at the 8th International Conference on Multimodal Interfaces (ICMI 2006), Banff, Alberta, Canada.
- Carlson, R. (1992). Synthesis: Modelling variability and constraints. *Speech Communication*, 11(2-3), 159-166.
- Carlson, R., Granström, B., & Karlsson, I. (1991). Experiments with voice modelling in speech synthesis. *Speech Communication*, 10(5), 481-489. doi:https://doi.org/10.1016/0167-6393(91)90051-T
- 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.
- Castellana, A., Carullo, A., Corbellini, S., & Astolfi, A. (2018). Discriminating Pathological Voice From Healthy Voice Using Cepstral Peak Prominence Smoothed Distribution in Sustained Vowel. *IEEE Transactions on Instrumentation and Measurement*, 67(3), 646-654. doi:10.1109/TIM.2017.2781958
- Chasaide, A., Ní Chiaráin, N., Wendler, C., Berthelsen, H., Murphy, A., & Gobl, C. (2017). *The ABAIR initiative: Bringing Spoken Irish into the Digital Space*. Paper presented at INTERSPEECH 2017, Stockholm, Sweden.
- Cheang, H. S., & Pell, M. D. (2009). Acoustic markers of sarcasm in Cantonese and English. *The Journal of the Acoustical Society of America*, 126(3), 1394-1405.
- Chen, G., Shue, Y.-L., Kreiman, J., & Alwan, A. (2012). *Estimating the voice source in noise*. Paper presented at INTERSPEECH 2012, Portland, OR.
- Choi, H., & Hahn, M. (2021). Sequence-to-Sequence Emotional Voice Conversion With Strength Control. *IEEE Access*, 9, 42674-42687. doi:10.1109/ACCESS.2021.3065460

- Chong, C. S., Kim, J., & Davis, C. (2018). Disgust expressive speech: The acoustic consequences of the facial expression of emotion. *Speech Communication, 98*, 68-72. doi:<https://doi.org/10.1016/j.specom.2017.12.007>
- Christensen, R. (2023). ordinal—Regression Models for Ordinal Data (Version R package version 2023.12-4.1). Retrieved from <https://CRAN.R-project.org/package=ordinal>
- Chronaki, G., Hardwin, J. A., Garner, M., Maurage, P., & Sonuga-Barke, E. J. S. (2014). The development of emotion recognition from facial expressions and non-linguistic vocalizations during childhood. *British Journal of Developmental Psychology, 33*, 218-236.
- Clore, G. L., & Ortony, A. (1991). What More Is There to Emotion Concepts Than Prototypes? *Journal of Personality and Social Psychology, 60*(1), 48-50. doi:<https://doi.org/10.1037/0022-3514.60.1.48>
- Cohen, A., & 't Hart, J. (1982). Declination: Construct or Intrinsic Feature of Speech? . *Phonetica, 39*(4/5), 254-273. doi:<https://doi.org/10.1159/000261666>
- Collignon, O., Girard, S., Gosselin, F., Roy, S., Saint-Amour, D., Lassonde, M., & Lepore, F. (2008). Audio-visual integration of emotion expression. *Brain Research, 1242*, 126-135.
- Collignon, O., Girard, S., Gosselin, F., Saint-Amour, D., Lepore, F., & Lassonde, M. (2010). Women process multisensory emotion expressions more efficiently than men. *Neuropsychologia, 48*(1), 220-225. doi:<https://doi.org/10.1016/j.neuropsychologia.2009.09.007>
- Conroy, D., Vitale, T., & Klatt, D. H. (1986). DECTalk DTC03 Text-to-Speech System Owner's Manual (Version EK-DTC03-OM-001). Nashua, NH: Educational Services of Digital Equipment Corporation.
- Cordaro, D. C., Keltner, D., Tshering, S., Wangchuk, D., & Flynn, L. M. (2016). The voice conveys emotion in ten globalized cultures and one remote village in Bhutan. *Emotion, 16*(1).
- Cordaro, D. C., Sun, R., Keltner, D., Kamble, S., Huddar, N., & McNeil, G. (2018). Universals and Cultural Variations in 22 Emotional Expressions Across Five Cultures. *Emotion, 18*(1), 75-93. doi:[10.1037/emo0000302](https://doi.org/10.1037/emo0000302)
- Cornelius, R. R. (2000). *Theoretical approaches to emotion*. Paper presented at the ITRW on Speech and Emotion, Newcastle, Northern Ireland.
- Cornille, T., Wang, F., & Bekker, J. (2022, 23-27 May 2022). *Interactive Multi-Level Prosody Control for Expressive Speech Synthesis*. Paper presented at the ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
- Cowen, A. S., Elfenbein, H. A., Laukka, P., & Keltner, D. (2019). Mapping 24 Emotions Conveyed by Brief Human Vocalization. *American Psychologist, 74*(6), 698-712.
- Cowen, A. S., Laukka, P., Elfenbein, H. A., Liu, R., & Keltner, D. (2019). The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nature Human Behaviour, 3*(4), 369-382. doi:[10.1038/s41562-019-0533-6](https://doi.org/10.1038/s41562-019-0533-6)
- 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. 109-114): World Scientific Engineering Society.
- Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. (2001). Emotion Recognition in Human-Computer Interaction. *IEEE Signal Processing Magazine*, 18(1), 32-80.
- Cranen, B., & Boves, L. (1985). Pressure measurements during speech production using semiconductor miniature pressure transducers: impact on models for speech production. *J Acoust Soc Am*, 77(4), 1543-1551. doi:10.1121/1.391997
- Creer, S., Cunningham, S., Green, P., & Yamagishi, J. (2013). Building personalised synthetic voices for individuals with severe speech impairment. *Computer Speech & Language*, 27, 1178-1193.
- Crivelli, C., Jarillo, S., Russell, J., & Fernandez-Dols, J.-M. (2016). Reading Emotions From Faces in Two Indigenous Societies. *Journal of experimental psychology. General*, 145(7). doi:10.1037/xge0000172
- Crumpton, J., & Bethel, C. (2015). A Survey of Using Vocal Prosody to Convey Emotion in Robot Speech. *International Journal of Social Robotics*, 8, 271-285. doi:10.1007/s12369-015-0329-4
- Cruttenden, A. (1997). *Intonation* (2nd ed.). Cambridge: Cambridge University Press.
- Cummings, K. E., & Clements, M. A. (1995). Glottal models for digital speech processing: a historical survey and new results. *Digital Signal Processing*, 5, 21-42.
- Dalton, J., Kane, J., Yanushevskaya, I., Ní Chasaide, A., & Gobl, C. (2014). *GloRí - the Glottal Research Instrument*. Paper presented at Speech Prosody 2014, Dublin, Ireland.
- Dalton, M. (2008). *The Phonetics and Phonology of the Intonation of Irish Dialects*. (PhD Thesis). Trinity College Dublin.
- Dalton, M., & Ní Chasaide, A. (2003). *Modelling Intonation in Three Irish Dialects*. Paper presented at the 15th International Congress of Phonetic Sciences (ICPhS XV), Barcelona, Spain.
- Dalton, M., & Ní Chasaide, A. (2005). Tonal Alignment in Irish Dialects. *Language and Speech*, 48(4), 441-464. doi:10.1177/00238309050480040501
- Dalton, M., & Ní Chasaide, A. (2007). *Nuclear accents in four Irish (Gaelic) dialects*. Paper presented at the 16th International Congress of Phonetic Sciences (ICPhS XVI), Saarbrücken, Germany.
- 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. (1998). *The expression of the emotions in man and animals* (3rd ed.). London: Harper-Collins.
- Dattorro, J. (1997). Effect Design, Part 2: Delay Line Modulation and Chorus. *Journal of The Audio Engineering Society*, 45, 764-788.

- Daz Productions, I. (2022). Daz Studio 4.22 (Version 4.22.0.16 (64-bit)) [Computer Software]. <https://www.daz3d.com/>
- de Gelder, B. (2000). Recognizing emotions by ear and by eye. In R. D. Lane & L. Nadel (Eds.), *Cognitive Neuroscience of Emotion* (pp. 84-105). New York: Oxford University Press.
- de Gelder, B., & Vroomen, J. (2000). The perception of emotion by ear and by eye. *Cognition & Emotion*, 14(3), 289-311. doi:10.1080/026999300378824
- Dehghanzadeh, H., Fardanesh, H., Hatami, J., Talaei, E., & Noroozi, O. (2019). Using gamification to support learning English as a second language: a systematic review. *Computer Assisted Language Learning*, 1-24. doi:10.1080/09588221.2019.1648298
- Department of Tourism, C., Arts, Gaeltacht, Sport and Media. (2023). *Digital Plan for the Irish Language: Speech and Language Technologies 2023-2027*. Government of Ireland.
- Ding, H., & Zhang, Y. (2023). Speech Prosody in Mental Disorders. *Annual Review of Linguistics*, 9, 335-355.
- Dorn, A., & Ní Chasaide, A. (2016). *Donegal Irish rises – similarities and differences to rises in English varieties*. Paper presented at Speech Prosody 2016, Boston, MA, USA.
- Dorn, A., O'Reilly, M., & Ní Chasaide, A. (2011). *Prosodic signalling of sentence mode in two varieties of Irish (Gaelic)*. Paper presented at the 15th International Congress of Phonetic Sciences (ICPhS XV), Hong Kong, China.
- Dorn, A., O'Reilly, M., & Ní Chasaide, A. (2012). *A preliminary analysis of speech rhythm in three varieties of Irish (Gaelic)*. Paper presented at Speech Prosody 2012, Shanghai, China.
- Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P. (2003). Emotional speech: towards a new generation of databases. *Speech Communication*, 40(1-2), 33-60. doi:[https://doi.org/10.1016/S0167-6393\(02\)00070-5](https://doi.org/10.1016/S0167-6393(02)00070-5)
- 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 - Workshop on Corpora for Research on Emotion and Affect, Marrakesh, Morocco.
- Drioli, C., Tisato, G., Cosi, P., & Tesser, F. (2003). *Emotions and voice quality: experiments with sinusoidal modelling*. Paper presented at the VOQUAL'03, Geneva, Switzerland.
- Dubagunta, S. P., Vlasenko, B., & Magimai-Doss, M. (2019). *Learning voice source related information for depression detection*. Paper presented at the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, England.
- Eibl-Eibesfeldt, I. (1989). *Human ethology*. New York, NY: Aldine.
- Ekman, P. (1982). *Emotion in the Human Face*. Cambridge: Cambridge University Press.
- Ekman, P. (1989). The argument and evidence about universals in facial expressions of emotion. In *Handbook of social psychophysiology*. (pp. 143-164). Oxford, England: John Wiley & Sons.
- Ekman, P. (1992a). Are there basic emotions? *Psychological Review*, 99(3), 550-553.

- Ekman, P. (1992b). An argument for basic emotions. *Cognition & Emotion*, 6, 169-200.
- Ekman, P. (1994). Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. *Psychological Bulletin*, 115(2), 268-287. doi:<https://doi.org/10.1037/0033-2909.115.2.268>
- Ekman, P. (1999). Basic emotions. In *Handbook of cognition and emotion*. (pp. 45-60). New York, NY, US: John Wiley & Sons Ltd.
- Ekman, P., & Cordaro, D. C. (2011). What is meant by calling emotions basic. *Emotion Review*, 3(4), 364-370.
- Ekman, P., & Friesen, W. V. (1975). *Pictures of facial affect*. Palo Alto, CA: Consulting Psychologists Press.
- Ekman, P., Friesen, W. V., & Hager, J. C. (2002). *Facial Action Coding System: The Manual On CD ROM*. Salt Lake City, UT: Research Nexus.
- Ekman, P., Levenson, R. W., & Friesen, W. V. (1983). Autonomic nervous system activity distinguishes among emotions. *Science*, 221(4616), 1208-1210. doi:<https://doi.org/10.1126/science.6612338>
- Ekman, P., Sorenson, E. R., & Friesen, W. V. (1969). Pan-Cultural Elements in Facial Displays of Emotion. *Science*, 164, 86-88.
- Elfenbein, H. A. (2013). Nonverbal Dialects and Accents in Facial Expressions of Emotion. *Emotion Review*, 5, 90. doi:10.1177/1754073912451332
- Elfenbein, H. A., & Ambady, N. (2002). On the Universality and Cultural Specificity of Emotion Recognition: A Meta-Analysis. *Psychological Bulletin*, 128(2), 203-235. doi:10.1037/0033-2909.128.2.203
- Elfenbein, H. A., Beaupré, M., Lévesque, M., & Hess, U. (2007). Toward a dialect theory: Cultural differences in the expression and recognition of posed facial expressions. *Emotion*, 7(1), 131-146. doi:10.1037/1528-3542.7.1.131
- Elfenbein, H. A., Mandal, M. K., Ambady, N., Harizuka, S., & Kumar, S. (2002). Cross-cultural patterns in emotion recognition: Highlighting design and analytical techniques. *Emotion*, 2(1), 75-84. doi:<https://doi.org/10.1037/1528-3542.2.1.75>
- Epstein, M. A. (2003). *Epstein, M. (2003). Voice quality and prosody in English*. Paper presented at the 15th International Congress of Phonetic Sciences (ICPhS XV), Barcelona, Spain.
- Erro, D., Moreno, A., & Bonafonte, A. (2010). INCA Algorithm for Training Voice Conversion Systems From Nonparallel Corpora. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(5), 944-953. doi:10.1109/TASL.2009.2038669
- Eskritt, M., & Kang, L. (2003). Do Actions Speak Louder Than Words? Preschool Children's Use of the Verbal-Nonverbal Consistency Principle During Inconsistent Communications. *Journal of Nonverbal Behavior*, 27(1), 25-41.
- Esling, J. H. (1978). The identification of features of voice quality in social groups. *Journal of the International Phonetic Association*, 8(1-2), 18-23. doi:10.1017/S0025100300001699

- Esling, J. H., Moisiuk, S. R., Benner, A., & Crevier-Buchman, L. (2019). *Voice Quality: The Laryngeal Articulator Model*. Cambridge: Cambridge University Press.
- 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, Speech, Music and Hearing, Royal Institute of Technology, Stockholm*, 36(2-3), 119-156.
- Fant, G. (1997). The voice source in connected speech. *Speech Communication*, 22, 125-139.
- Fant, G., Liljencrants, J., & Lin, Q.-g. (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.
- Fernandez, R., & Ramabhadran, B. (2007). *Automatic Exploration of Corpus-Specific Properties for Expressive Text-to-Speech: A Case Study in Emphasis*. Paper presented at the 6th ISCA Workshop on Speech Synthesis, Bonn, Germany.
- Fitch, W. T., & Giedd, J. (1999). Morphology and development of the human vocal tract: a study using magnetic resonance imaging. *Journal of the Acoustical Society of America*, 106 (3 Pt 1), 1511-1522. doi:10.1121/1.427148
- 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. doi:https://doi.org/10.1037/0022-3514.65.5.942
- Fónagy, I. (1978). A new method of investigating the perception of prosodic features. *Language & Speech*, 21(1), 34-49. doi:10.1177/002383097802100102
- Fónagy, I. (2001). *Languages within Language: an Evolutive Approach*. Amsterdam, Netherlands: John Benjamins Publishing Company.
- Fónagy, I., & Magdics, K. (1963). Emotional patterns in intonation and music. *Zeitschrift für Phonetik*(16), 293-326.
- Fourier, J. B. J. (1878). *The Analytical Theory of Heat* (A. Freeman, Trans.). London, United Kingdom: The University Press.
- Frick, R. W. (1985). Communicating Emotion: The Role of Prosodic Features. *Psychological Bulletin*, 97(3), 412-429.
- Friend, M., & Farrar, M. J. (1994). A comparison of content-masking procedures for obtaining judgments of discrete affective states. *Journal of The Acoustical Society of America*, 96(3), 1283-1290. doi:https://doi.org/10.1121/1.410276
- Fröhlich, M., Michaelis, D., & Strube, H. W. (2001). SIM—simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals. *The Journal of the Acoustical Society of America*, 110(1), 479-488. doi:10.1121/1.1379076

- 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), Tokyo, Japan.
- Garellek, M. (2014). Voice quality strengthening and glottalization. *Journal of Phonetics*, 45, 106-113. doi:<https://doi.org/10.1016/j.wocn.2014.04.001>
- Garellek, M. (2022). Theoretical achievements of phonetics in the 21st century: Phonetics of voice quality. *Journal of Phonetics*, 94, 101155. doi:<https://doi.org/10.1016/j.wocn.2022.101155>
- Gendron, M., Roberson, D., van der Vyver, J. M., & Barrett, L. F. (2014a). Cultural Relativity in Perceiving Emotion from Vocalizations. *Psychological Science*, 25(4), 911-920.
- Gendron, M., Roberson, D., van der Vyver, J. M., & Barrett, L. F. (2014b). Perceptions of Emotion From Facial Expressions Are Not Culturally Universal: Evidence From a Remote Culture. *Emotion*, 14, 251-262.
- Gobl, C. (2006). *Modelling aspiration noise during phonation using the LF voice source model*. Paper presented at the 8th International Conference on Spoken Language Processing, INTERSPEECH 2006, Pittsburgh.
- Gobl, C. (2021). *The LF Model in the Frequency Domain for Glottal Airflow Modelling without Aliasing Distortion*. Paper presented at INTERSPEECH 2021, Brno, Czech Republic.
- Gobl, C., Bennett, E., & Ní Chasaide, A. (2002). Expressive synthesis: how crucial is voice quality? In (pp. 91-94). Piscataway, NJ, USA, USA: IEEE.
- Gobl, C., & Ní Chasaide, A. (1999). Techniques for analysing the voice source. In W. J. Hardcastle & N. Hewlett (Eds.), *Coarticulation: Theory, Data and Techniques* (pp. 300-320). 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. (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), 189-212. doi:10.1016/S0167-6393(02)00082-1
- Gobl, C., & Ní Chasaide, A. (2010). Voice Source Variation and Its Communicative Functions. In W.J. Hardcastle, J. Laver & F.E. Gibbon (Eds.), *The Handbook of Phonetic Sciences* (pp. 378-423). Wiley-Blackwell.
- Gong, C., Wang, L., Ling, Z., Guo, S., Zhang, J., & Dang, J. (2021). *Improving Naturalness and Controllability of Sequence-to-Sequence Speech Synthesis by Learning Local Prosody Representations*. Paper presented at the ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada.

- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, 27, 2672--2680.
- Gordon, M., & Ladefoged, P. (2001). Phonation types: a cross-linguistic overview. *Journal of Phonetics*, 29(4), 383-406. doi:<https://doi.org/10.1006/jpho.2001.0147>
- Govind, D., & Prasanna, S. R. M. (2013). Expressive speech synthesis: a review. *International Journal of Speech Technology*, 16, 237-260. doi:10.1007/s10772-012-9180-2
- Grabe, E. (1998). *Comparative intonational phonology: English and German*. (PhD Thesis). Max-Plank Institute and Radboud University of Nijmegen, Nijmegen.
- Grabe, E. (2001). The IViE Labelling Guide Version 3. <http://www.phon.ox.ac.uk/files/apps/IViE/guide.html>
- Grabe, E., Nolan, F., & Farrar, K. J. (1998). *IViE - A Comparative Transcription system for Intonational Variation in English*. Paper presented at the 5th International Conference on Spoken Language Processing (ICSLP 98), Sydney, Australia.
- Grandjean, D., Bänziger, T., & Scherer, K. R. (2006). Intonation as an interface between language and affect. *Progress on Brain Research*, 156, 235-268.
- Granqvist, S. (1996). Granqvist, S. (1996). Enhancements to the visual analogue scale, VAS, for listening tests. *TMH-QPSR*, 37(4), 61-65.
- Greasley, P., Sherrard, C., & Waterman, M. (2000). Emotion in language and speech: Methodological issues in naturalistic approaches. *Language and Speech*, 43(4), 355-375. doi:<https://doi.org/10.1177/00238309000430040201>
- Grichkovtsova, I., Morel, M., & Lacheret, A. (2012). The role of voice quality and prosodic contour in affective speech perception. *Speech Communication*, 54(3), 414-429. doi:<https://doi.org/10.1016/j.specom.2011.10.005>
- Guerrero, L. K., & Andersen, P. A. (1998). Jealousy Experience and Expression in Romantic Relationships. In P. A. Andersen & L. K. Guerrero (Eds.), *Handbook of communication and emotion: Research, Theory, Applications, and Contexts* (pp. 155-188). New York: Academic Press.
- Guerrero, L. K., Andersen, P. A., & Trost, M. R. (1998). Communication and Emotion: Basic Concepts and Approaches. In P. A. Andersen & L. K. Guerrero (Eds.), *Handbook of communication and emotion: Research, Theory, Applications, and Contexts* (pp. 3-27). New York: Academic Press.
- Gussenhoven, C. (1984). *On the grammar and semantics of sentence accents*. Dordrecht: Foris.
- Gussenhoven, C., Repp, B. G., Rietveld, A., Rump, H. H., & Terken, J. (1997). The perceptual prominence of fundamental frequency peaks. *The Journal of the Acoustical Society of America*, 102(5 Pt. 1), 3009-3022.
- Guzman, M., Correa, S., Muñoz, D., & Mayerhoff, R. (2013). Influence on Spectral Energy Distribution of Emotional Expression. *Journal of Voice*, 27(1), 129.E121-129.E110. doi:<https://doi.org/10.1016/j.jvoice.2012.08.008>

- Haan, J., Van Heuven, V. J., Pacilly, J. J. A., & van Bezooijen, R. (1997). An anatomy of Dutch question intonation. *Linguistics in the Netherlands*, 14, 97-108.
- Haidt, J., & Keltner, D. (1999). Culture and facial expression: Open-ended methods find more expressions and a gradient of recognition. *Cognition & Emotion*, 13, 225-266.
- Hall, J. A., Coats, E. J., & LeBeau, L. S. (2005). Nonverbal Behavior and the Vertical Dimension of Social Relations: A Meta-Analysis. *Psychological Bulletin*, 131, 898-924.
- Hall, K. (1995). Lip Service on the Fantasy Lines. In M. Bucholtz & K. Hall (Eds.), *Gender articulated: Language and the socially constructed self*. New York, NY: Routledge.
- Hao, S., Liang, L., Wang, J., Liu, H., & Chen, B. (2021). The effects of emotional context and exposure frequency on L2 contextual word learning. *International Journal of Bilingualism*, 25(5), 1280-1296. doi:10.1177/13670069211016540
- Harrigan, J. A., Lucic, K. S., & Rosenthal, R. (1991). Retelling anxious events: Effects on trait and state anxiety. *Personality and Individual Difference*, 12(9), 917-927. doi:https://doi.org/10.1016/0191-8869(91)90181-A
- Heinrich, N., d'Alessandro, C., & Doval, B. (2001). *Spectral correlates of voice open quotient and glottal flow asymmetry: theory, limits and experimental data*. Paper presented at the 7th European Conference on Speech Communication and Technology (Eurospeech 2001), Scandinavia.
- Hejrná, M., Šturm, P., Tylečková, L., & Bořil, T. (2021). Normophonic Breathiness in Czech and Danish: Are Females Breathier Than Males? *Journal of Voice*, 35(3), 498.e491-498.e422. doi:https://doi.org/10.1016/j.jvoice.2019.10.019
- Henton, C. G., & Bladon, A. (1988). Creak as a sociophonetic marker. In L. Hyman, V. Fromkin, & C. N. Li (Eds.), *Language, speech and mind: Studies in honour of Victoria A. Fromkin* (pp. 3-29). London, UK: Routledge.
- Henton, C. G., & Bladon, R. A. W. (1985). Breathiness in normal female speech: Inefficiency versus desirability. *Language & Communication*, 5(3), 221-227. doi:https://doi.org/10.1016/0271-5309(85)90012-6
- Hermes, D. J. (2006). Stylization of pitch contours. In S. Sudhoff & D. Lenertova (Eds.), *Methods in empirical prosody research* (pp. 29-62). Berlin, New York: Walter de Gruyter GmbH.
- Heuft, B., Portele, T., & Rauth, M. (1996). Emotions in time domain synthesis. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, 3, 1974-1977 vol.1973.
- Hickey, R. (2011). *The Dialects of Irish: study of a changing landscape*. Berlin, Boston: De Gruyter Mouton.
- Hjelmquist, E., Dahlstrand, U., & Hedelin, L. (1992). Visually Impaired Person's Comprehension of text Presented with Speech Synthesis. *Journal of Visual Impairment & Blindness*, 86(10), 426-428. doi:https://doi.org/10.1177/0145482X9208601005
- Hollien, H. (1974). On vocal registers. *Journal of Phonetics*, 2, 125-143.

- Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2), 65-70. Retrieved from <http://www.jstor.org/stable/4615733>
- Holmes, J. (1973). The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer. *IEEE Transactions on Audio and Electroacoustics*, 21(3), 298-305. doi:10.1109/TAU.1973.1162466
- Hono, Y., Tsuboi, K., Sawada, K., Hashimoto, K., Oura, K., Nankaku, Y., & Tokuda, K. (2020). *Hierarchical Multi-Grained Generative Model for Expressive Speech Synthesis*. Paper presented at INTERSPEECH 2020, Shanghai, China.
- Hunt, A. J., & Black, A. W. (1996). *Unit selection in a concatenative speech synthesis system using a large speech database*. Paper presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing Conference, Atlanta, GA, USA.
- Huttar, G. L. (1968). Relations between prosodic variables and emotions in normal American English utterances. *Journal of Speech & Hearing Research*, 11(3), 481-487. doi:<https://doi.org/10.1044/jshr.1103.481>
- Iida, A., Campbell, N., Iga, S., Higuchi, F., & Yasumura, M. (2000). *A speech synthesis system with emotion for assisting communication*. Paper presented at the ISCA Workshop on Speech & Emotion, Northern Ireland.
- Iida, A., Campbell, N., & Yasumura, M. (1998). *Emotional speech as an effective interface for people with special needs*. Paper presented at the 3rd Asia Pacific Computer Human Interaction (Cat. No.98EX110).
- Ilie, G., & Thompson, W. F. (2006). A Comparison of Acoustic Cues in Music and Speech for Three Dimensions of Affect. *Music Perception*, 23(4), 319-329. doi:10.1525/mp.2006.23.4.319
- Im, C.-B., Lee, S.-H., Kim, S.-B., & Lee, S.-W. (2022, 23-27 May 2022). *EMOQ-TTS: Emotion Intensity Quantization for Fine-Grained Controllable Emotional Text-to-Speech*. Paper presented at the ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore.
- Inoue, K., Hara, S., Abe, M., Hojo, N., & Iijima, Y. (2017). *An investigation to transplant emotional expressions in DNN-based TTS synthesis*. Paper presented at the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, Malaysia.
- Inoue, K., Hara, S., Abe, M., Hojo, N., & Iijima, Y. (2021). Model architectures to extrapolate emotional expressions in DNN-based text-to-speech. *Speech Communication*, 126, 35-43. doi:<https://doi.org/10.1016/j.specom.2020.11.004>
- 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 INTERSPEECH 2006 - ICSLP, Pittsburgh, Pennsylvania, USA.
- Ishii, K., Reyes, J. A., & Kitayama, S. (2003). Spontaneous attention to word content versus emotional tone: Differences among three cultures. *Psychological Science*, 14(1), 39-46.

- Ishizaka, K., & Flanagan, J. L. (1972). Acoustic Properties of a Two-Mass Model of the Vocal Cords. *The Journal of the Acoustical Society of America*, 51(1A_Supplement), 91. doi:10.1121/1.1981696
- Ishizaka, K., & Matsudaira, M. (1968). *What makes the vocal cords vibrate?* Paper presented at the 6th International Congress on Acoustics, Tokyo, Japan.
- Izard, C. E. (1990). Facial Expressions and the regulation of emotions. *Journal of Personality and Social Psychology*, 58(3), 487-498.
- Izard, C. E. (1993). Four systems for emotion activation: Cognitive and noncognitive processes. *Psychological Review*, 100(1), 68-90. doi:10.1037/0033-295X.100.1.68
- Izard, C. E. (1994). Innate and universal facial expressions: evidence from developmental and cross-cultural research. *Psychological Bulletin*, 115(2), 288-299.
- Jack, R. E., Blais, C., Scheepers, C., Schyns, P. G., & Caldara, R. (2009). Cultural Confusions Show that Facial Expressions Are Not Universal. *Current Biology*, 19(18), 1543-1548. doi:https://doi.org/10.1016/j.cub.2009.07.051
- Jack, R. E., Garrod, O. G. B., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, 109(19), 7241-7244. doi:https://doi.org/10.1073/pnas.1200155109
- James, W. (1884). What is emotion? *Mind*, 19, 188-205.
- Jia, J., Zhang, S., Meng, F., Wang, Y., & Cai, L. (2011). Emotional Audio-Visual Speech Synthesis Based on PAD. *IEEE Transactions on Audio, Speech, and Language Processing*, 19, 570-582.
- Jiang, X., & Pell, M. D. (2017). The sound of confidence and doubt. *Speech Communication*, 88, 106-126. doi:https://doi.org/10.1016/j.specom.2017.01.011
- Johnson-Laird, P. N., & Oatley, K. (1989). The language of emotions: an analysis of a semantic field. *Cognition & Emotion*, 3(2), 81-123.
- Johnson, K. (2011). *Acoustic and Auditory Phonetics* (3 ed.). Hoboken, NJ: Wiley-Blackwell
- 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 (ICPhS XIII), Stockholm, Sweden.
- Johnstone, T., & Scherer, K. R. (1999). *The effects of emotions on voice quality*. Paper presented at the 14th International Congress of Phonetic Sciences (ICPhS XIV), San Francisco, USA.
- Johnstone, T., & Scherer, K. R. (2000). Vocal communication of Emotion. In M. Lewis & J. M. Haviland-Jones (Eds.), *Handbook of Emotions* (pp. 220-235). New York: Guilford Press.
- Jones, D. (1960). *An Outline of English Phonetics* (9 ed.). Cambridge: Heffer.
- Jürgens, R., Hammerschmidt, K., & Fischer, J. (2011). Authentic and play-acted vocal emotion expressions reveal acoustic differences. *Frontiers in Psychology*, 2, 180.

- Juslin, P. N., Harmat, L., & Laukka, P. (2018). The wisdom of the body: Listeners' autonomic arousal distinguishes between spontaneous and posed vocal emotions. *Scandinavian Journal of Psychology*, 59, 105-112.
- 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*, 129(5), 770-814. doi:10.1037/0033-2909.129.5.770
- Juslin, P. N., Laukka, P., & Bänziger, T. (2018). The mirror to our soul? A comparison of spontaneous and posed vocal expression of emotion. *Journal of Nonverbal Behavior*, 42, 1-40.
- Juslin, P. N., Laukka, P., Harmat, L., & Ovsianikow, M. (2021). Spontaneous vocal expressions from everyday life convey discrete emotions to listeners. *Emotion*, 21(6), 1281-1301.
- 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., & Gobl, C. (2013). Automating manual user strategies for precise voice source analysis. *Speech Communication*, 55, 397-414.
- Kane, J., Yanushevskaya, I., Ní Chasaide, A., & Gobl, C. (2012). *Exploiting Time and Frequency Domain Measures for Precise Voice Source Parameterisation*. Paper presented at Speech Prosody 2012, Shanghai, China.
- Kapitanović, A., Tokić, A., & Šimić, N. (2022). Differences in the recognition of sadness, anger, and fear in facial expressions: the role of the observer and model gender. *Arh Hig Rada Toksikol*, 73(4), 308-313. doi:10.2478/aiht-2022-73-3662
- Kappas, A., Hess, U., & Scherer, K. (1991). Voice and Emotion. In R. S. Feldman & B. Rimé (Eds.), *Fundamentals of Non-verbal Behavior* (pp. 200-238). Cambridge: Cambridge University Press.
- Kasuya, H., Maekawa, K., & Kiritani, S. (1999). *Joint estimation of voice source and vocal tract parameters as applied to the study of voice source dynamics*. Paper presented at the 14th International Congress of Phonetic Sciences (ICPhS XIV), San Francisco, CA.
- Kawahara, H. (1997). *Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited*. Paper presented at the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, Munich, Germany, pp. 1303-1306 vol.2. doi: 10.1109/ICASSP.1997.596185
- Kawahara, H. (2006). STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. *Acoustical Science and Technology*, 27(6), 349-353. doi:10.1250/ast.27.349
- Kawahara, H., Masuda-Katsuse, I., & de Cheveigné, A. (1999). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, 27(3), 187-207. doi:https://doi.org/10.1016/S0167-6393(98)00085-5

- Keltner, D., & Ekman, P. (2000). Facial Expression of Emotion. In M. Lewis & J. M. Haviland-Jones (Eds.), *Handbook of Emotions* (pp. 236-249). New York: Guilford Press.
- Keltner, D., & Haidt, J. (1999). Social functions of emotions at four levels of analysis. *Cognition & Emotion*, 13(5), 505-521.
- Keltner, D., Oatley, K., & Jenkins, J. M. (2019). *Understanding Emotions* (4 ed.): John Wiley & Sons.
- Keltner, D., Sauter, D., Tracy, J., & Cowen, A. (2019). Emotional Expression: Advances in Basic Emotion Theory. *Journal of Nonverbal Behavior*, 43(2), 133-160.
- Keltner, D., Tracy, J., Sauter, D. A., Cordaro, D. C., & McNeil, G. (2016). Expression of Emotion. In L. F. Barrett & M. Lewis (Eds.), *Handbook of Emotions* (4 ed., pp. 467-482). New York, NY: Guilford Press.
- Kim, C., Um, S., Yoon, H., & Kang, H.-G. (2022). *FluentTTS: Text-dependent Fine-grained Style Control for Multi-style TTS*. Paper presented at INTERSPEECH 2022, Incheon, Korea.
- Kim, S. K., & Sumner, M. (2017). Beyond lexical meaning: The effect of emotional prosody on spoken word recognition. *Journal of The Acoustical Society of America*, 142(1). doi:<https://doi.org/10.1121/1.4991328>
- King, S. (2011). An introduction to statistical parametric speech synthesis. *Sadhana*, 36, 837-852.
- Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. *ICLR 2014*. doi: <https://doi.org/10.48550/arXiv.1312.6114>
- Kitayama, S., & Ishii, K. (2002). Word and voice: Spontaneous attention to emotional utterances in two languages. *Cognition & Emotion*, 16(1), 29-59.
- Kitayama, S., Mesquita, B., & Karasawa, M. (2006). Cultural affordances and emotional experience: Socially engaging and disengaging emotions in Japan and the United States. *Journal of Personality and Social Psychology*, 91(5), 890-903. doi:<https://doi.org/10.1037/0022-3514.91.5.890>
- Kitzing, P., & Löfqvist, A. (1975). Subglottal and oral air pressures during phonation—preliminary investigation using a miniature transducer system. *Medical and biological engineering*, 13(5), 644-648. doi:10.1007/BF02477320
- Klasmeyer, G. (1997). *The perceptual importance of selected voice quality parameters*. Paper presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, Munich, Germany.
- 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. *Journal of The Acoustical Society of America*, 67(3), 971-995.
- Klatt, D. H. (1987). Review of text-to-speech conversion for English. *Journal of The Acoustical Society of America*, 82(3), 737-793.

- Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of The Acoustical Society of America*, 87(2), 820-857.
- Knight, R.-A. (2008). The Shape of Nuclear Falls and their Effect on the Perception of Pitch and Prominence: Peaks vs. Plateaux. *Language and Speech*, 51(3), 223-244.
doi:<https://doi.org/10.1177/0023830908098541>
- Knowles, G. (1978). The nature of phonological variables in Scouse. In P. Trudgill (Ed.), *Sociolinguistic patterns in British English*. London, UK: Edward Arnold.
- Koeda, M., Belin, P., Hama, T., Masuda, T., Matsuura, M., & Okubo, Y. (2013). Cross-Cultural Differences in the Processing of Non-Verbal Affective Vocalizations by Japanese and Canadian Listeners. *Frontiers in Psychology*, 4, 105. doi:10.3389/fpsyg.2013.00105
- Kotz, S. A., & Paulmann, S. (2007). When emotional prosody and semantics dance cheek to cheek: ERP evidence. *Brain Research*, 1151, 107-118.
- Kraus, M. W. (2017). Voice-only communication enhances empathic accuracy. *American Psychologist*, 72, 644-654.
- Kraus, M. W., Park, J. W., & Tan, J. J. X. (2017). Signs of social class: The experience of economic inequality in everyday life. *Perspectives on Psychological Science*, 12(3), 422-435.
- Kreiman, J., Garellek, M., Chen, G., Alwan, A., & Gerratt, B. R. (2015). Perceptual evaluation of voice source models. *J Acoust Soc Am*, 138(1), 1-10. doi:10.1121/1.4922174
- Kreiman, J., & Sidtis, D. (2011). Perception of Emotion and Personality from Voice. In *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception* (pp. 302-360): Wiley-Blackwell.
- Ladd, D. R. (2008). *Intonational phonology* (2nd ed.). Cambridge, UK: Cambridge University Press.
- Ladd, D. R., Silverman, K., Tolkmitt, F. J., Bergmann, G., & Scherer, K. R. (1985). Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect. *Journal of The Acoustical Society of America*, 78, 435-444.
doi:10.1121/1.392466
- Ladefoged, P. (1971). *Preliminaries to linguistic phonetics*. Chicago: University of Chicago Press.
- Łańcucki, A. (2020). FastPitch: Parallel Text-to-speech with Pitch Prediction (Version 2). Retrieved from arXiv:2006.06873v2 [eess.AS]
- Larrouy-Maestri, P., Poeppel, D., & Pell, M. D. (2024). The Sound of Emotional Prosody: Nearly 3 Decades of Research and Future Directions. *Perspectives on Psychological Science*, 20(4), 623-638. doi:10.1177/17456916231217722
- 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. I, pp. 153-169). San Diego, CA: Plural Publishing.
- Laukka, P., & Elfenbein, H. A. (2012). Emotion appraisal dimensions can be inferred from vocal expressions. *Social Psychological and Personality Science*, 3(5), 529-536.
doi:<https://doi.org/10.1177/1948550611428011>

- Laukka, P., Elfenbein, H. A., Söder, N., Nordström, H., Althoff, J., Chui, W., . . . Thingujam, N. S. (2013). Cross-cultural decoding of positive and negative non-linguistic emotion vocalizations. *Frontiers in Psychology*, 4.
- Laukka, P., Elfenbein, H. A., Thingujam, N., Rockstuhl, T., Iraki, F., Chui, W., & Althoff, J. (2016). The Expression and Recognition of Emotions in the Voice Across Five Nations: A Lens Model Analysis Based on Acoustic Features. *Journal of Personality and Social Psychology*, 111. doi:10.1037/pspi0000066
- Laukka, P., Neiberg, D., & Elfenbein, H. A. (2014). Evidence for Cultural Dialects in Vocal Emotion Expression: Acoustic Classification Within and Across Five Nations. *Emotion*, 14(3). doi:10.1037/a0036048
- 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* (Vol. I). 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(3), 313-335. doi:https://doi.org/10.1006/jpho.1996.0017
- Laver, J. (1980). *The phonetic description of voice quality*. Cambridge: Cambridge University Press.
- Lavner, Y., Gath, I., & Rosenhouse, J. (2000). The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. *Speech Communication*, 30(1), 9-26. doi:https://doi.org/10.1016/S0167-6393(99)00028-X
- Le, M., Vyas, A., Shi, B., Karrer, B., Sari, L., Moritz, R., . . . Hsu, W.-N. (2023). Voicebox: Text-Guided Multilingual Universal Speech Generation at Scale. doi:https://doi.org/10.48550/arXiv.2306.15687
- Lee, K., Park, K., & Kim, D. (2021). *STYLER: Style Factor Modeling with Rapidity and Robustness via Speech Decomposition for Expressive and Controllable Neural Text to Speech*. Paper presented at INTERSPEECH 2021, Brno, Czech Republic.
- Lee, Y., Keating, P., & Kreiman, J. (2019). Acoustic voice variation within and between speakers. *The Journal of the Acoustical Society of America*, 146(3), 1568–1579. doi:10.1121/1.5125134
- Lee, Y., & Kim, T. (2019, 12-17 May 2019). *Robust and Fine-grained Prosody Control of End-to-end Speech Synthesis*. Paper presented at the ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK.
- Lee, Y., Rabiee, A., & Lee, S.-Y. (2017). Emotional End-to-End Neural Speech Synthesizer. *ArXiv, abs/1711.05447*. doi:https://doi.org/10.48550/arXiv.1711.05447
- Lei, Y., Yang, S., Wang, X., & Xie, L. (2022). MsEmoTTS: Multi-Scale Emotion Transfer, Prediction, and Control for Emotional Speech Synthesis. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 30, 853–864. doi:10.1109/taslp.2022.3145293
- Li, T., Wang, X., Xie, Q., Wang, Z., & Xie, L. (2022). Cross-Speaker Emotion Disentangling and Transfer for End-to-End Speech Synthesis. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 30, 1448–1460. doi:10.1109/taslp.2022.3164181

- Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54(5), 358–368. doi:<https://doi.org/10.1037/h0044417>
- Lieberman, P. (1968). *Intonation, perception, and language* Cambridge, MA: MIT Press.
- Lieberman, P., & Michaels, S. B. (1962). Some aspects of fundamental frequency and envelope amplitudes as related to the emotional content of speech. *Journal of The Acoustical Society of America*, 34(7), 922–927. doi:<https://doi.org/10.1121/1.1918222>
- Lin, Y., Ding, H., & Zhang, Y. (2020). Prosody Dominates Over Semantics in Emotion Word Processing: Evidence From Cross-Channel and Cross-Modal Stroop Effects. *Journal of Speech, Language, and Hearing Research*, 63, 896-912.
- Liu, P., Rigoulot, S., & Pell, M. D. (2015a). Cultural differences in on-line sensitivity to emotional voices: comparing East and West. *Frontiers in Human Neuroscience, Volume 9 - 2015*. doi:[10.3389/fnhum.2015.00311](https://doi.org/10.3389/fnhum.2015.00311)
- Liu, P., Rigoulot, S., & Pell, M. D. (2015b). Culture modulates the brain response to human expressions of emotion: Electrophysiological evidence. *Neuropsychologia*, 67, 1-13. doi:<https://doi.org/10.1016/j.neuropsychologia.2014.11.034>
- Lorenzo-Trueba, J., Henter, G. E., Takaki, S., Yamagishi, J., Morino, Y., & Ochiai, Y. (2018). Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis. *Speech Communication*, 99, 135-143. doi:<https://doi.org/10.1016/j.specom.2018.03.002>
- Luo, Z., Chen, J., Takiguchi, T., & Arik, Y. (2017). *Emotional Voice Conversion with Adaptive Scales F0 Based on Wavelet Transform Using Limited Amount of Emotional Data*. Paper presented at INTERSPEECH 2017, Stockholm, Sweden.
- Marsh, A. A., Effenbein, H. A., & Ambady, N. (2003). Nonverbal "Accents": Cultural Differences in Facial Expressions of Emotion. *Psychological Science*, 14(4), 373-376.
- Massaro, D. W. (2000). *Multimodal emotion perception: Analogous to speech processes*. Paper presented at the ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion, Newcastle, Northern Ireland.
- Massaro, D. W., & Egan, P. B. (1996). Perceiving affect from the voice and the face. *Psychonomic Bulletin & Review*, 3(2), 215-221. doi:<https://doi.org/10.3758/BF03212421>
- Matoušek, J., Krňoul, Z., Campr, M., Zajíc, Z., Hanzlíček, Z., Grüber, M., & Kocurová, M. (2020). Speech and web-based technology to enhance education for pupils with visual impairment. *Journal on Multimodal User Interfaces*, 14(2), 219-230. doi:[10.1007/s12193-020-00323-1](https://doi.org/10.1007/s12193-020-00323-1)
- Matsumoto, D., Yoo, S. H., & Fontaine, J. (2008). Mapping expressive differences around the world: The relationship between emotional display rules and individualism versus collectivism. *Journal of Cross-Cultural Psychology*, 39, 55-74.
- McClure, E. B. (2000). A meta-analytic review of sex differences in facial expression processing and their development in infants, children, and adolescents. *Psychol Bull*, 126(3), 424-453. doi:[10.1037/0033-2909.126.3.424](https://doi.org/10.1037/0033-2909.126.3.424)

- McDermott, P. (2011). 'Irish isn't spoken here?' Language policy and planning in Ireland. *English Today*, 27, No. 2(106), 25-31.
- Mehta, D. D., Espinoza, V. M., Van Stan, J. H., Zaňartu, M., & Hillman, R. E. (2019). The difference between first and second harmonic amplitudes correlates between glottal airflow and neck-surface accelerometer signals during phonation. *The Journal of the Acoustical Society of America*, 145(5), EL386-EL392. doi:10.1121/1.5100909
- Mendoza-Denton, N. (2011). The Semiotic hitchhiker's guide to creaky voice: Circulation and gendered hardcore in a Chicana/o gang persona. *Journal of Linguistic Anthropology*, 21(2).
- Mennen, I., Scobbie, J. M., de Leeuw, E., Schaeffler, S., & Schaeffler, F. (2010). Measuring language-specific phonetic settings. *Second Language Research*, 26(1), 13-41.
- Mertl, J., Žáčková, E., & Řepová, B. (2018). Quality of life of patients after total laryngectomy: the struggle against stigmatization and social exclusion using speech synthesis. *Disability and Rehabilitation: Assistive Technology*, 13, 342-352. doi:http://dx.doi.org/10.1080/17483107.2017.1319428
- Mohammadi, S. H., & Kain, A. (2017). An overview of voice conversion systems. *Speech Communication*, 88, 65-82. doi:https://doi.org/10.1016/j.specom.2017.01.008
- Moisik, S. R., & Esling, J. H. (2011). *The 'whole larynx' approach to laryngeal features*. Paper presented at the 17th International Congress of Phonetic Sciences (ICPhS XVII), Hong Kong.
- Mokhtari, P., Story, B., Alku, P., & Ando, H. (2018). Estimation of the glottal flow from speech pressure signals: Evaluation of three variants of iterative adaptive inverse filtering using computational physical modelling of voice production. *Speech Communication*, 104, 24-38.
- Monahan, P. (1996). *Systems for Voice Source Analysis*. (PhD Thesis). University of Dublin, Trinity College Dublin, Ireland.
- Monzo, C., Alías, F., Sanz, 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, E., Clements, M. A., Peifer, J. W., & Weissner, L. (2008). Critical Analysis of the Impact of Glottal Features in the Classification of Clinical Depression in Speech. *IEEE Transactions on Biomedical Engineering*, 55, 96-107.
- Morise, M., Yokomori, F., & Ozawa, K. (2016). WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. *IEICE Transactions on Information and Systems*, E99.D(7), 1877-1884. doi:10.1587/transinf.2015EDP7457
- Morningstar, M., Dirks, M. A., & Huang, S. (2017). Vocal Cues Underlying Youth and Adult Portrayals of Socio-emotional Expressions. *Journal of Nonverbal Behavior*, 41, 155-183.
- Morton, J. B., & Trehub, S. (2001). Children's Understanding of Emotion in Speech. *Child Development*, 72, 834-843. doi:10.1111/1467-8624.00318
- Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5-6), 453-467.

- Moulines, E., & Laroche, J. (1995). Non-parametric techniques for pitch-scale and time-scale modification of speech. *Speech Communication*, 16(2), 175-205.
doi:[https://doi.org/10.1016/0167-6393\(94\)00054-E](https://doi.org/10.1016/0167-6393(94)00054-E)
- Mozziconacci, S. (1995). *Pitch variations and emotion in speech*. Paper presented at the 13th International Congress on Photonic Sciences (ICPhS XIII), Stockholm, Sweden.
- Mozziconacci, S. (1998). *Speech Variability and Emotion: Production and Perception*. (PhD Thesis). Technical University Eindhoven, Eindhoven.
- Mozziconacci, S. (2002). *Prosody and emotions*. Paper presented at Speech Prosody 2002, Aix-en-Provence, France.
- Murphy, A. (2021). *Controlling the voice quality dimension of prosody in synthetic speech using an acoustic glottal model* (PhD Thesis). Trinity College Dublin, Dublin, Ireland.
- Murphy, A., Lonergan, L., Qian, M., Berthelsen, H., Wendler, C., Ní Chiaráin, N., . . . Gobl, C. (2023). *Abair & Éist: a demonstration of speech technologies for Irish*. Paper presented at the 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023).
- Murphy, A., Yanushevskaya, I., Chasaide, A., & Gobl, C. (2021). *Integrating a Voice Analysis-Synthesis System with a TTS Framework for Controlling Affect and Speaker Identity*. Paper presented at the 32nd Irish Signals and Systems Conference (ISSC), Athlone, Ireland.
- Murphy, A., Yanushevskaya, I., Ní Chasaide, A., & Gobl, C. (2017). *Rd as a control parameter to explore affective correlates of the tense-lax continuum* Paper presented at INTERSPEECH 2017, Stockholm, Sweden.
- Murphy, A., Yanushevskaya, I., Ní Chasaide, A., & Gobl, C. (2018). *Voice Source Contribution to Prominence Perception: RD Implementation*. Paper presented at INTERSPEECH 2018, Hyderabad, India.
- Murphy, A., Yanushevskaya, I., Ní Chasaide, A., & Gobl, C. (2019). *The Role of Voice Quality in the Perception of Prominence in Synthetic Speech*. Paper presented at INTERSPEECH 2019, Graz, Austria.
- Murphy, A., Yanushevskaya, I., Ní Chasaide, A., & Gobl, C. (2020). *Testing the GlórCáil System in a Speaker and Affect Voice Transformation Task*. Paper presented at the 10th International Conference on Speech Prosody, Tokyo, Japan.
- Murphy, A., Yanushevskaya, I., Ní Chasaide, A., & Gobl, C. (2022). *Affect Expression: Global and Local Control of Voice Source Parameters*. Paper presented at Speech Prosody 2022, Lisbon, Portugal.
- Murray, I. R. (1989). *Simulating emotion in synthetic speech*. (PhD Thesis). University of Dundee, United Kingdom.
- 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. (1995). Implementation and testing of a system for producing emotion-by-rule in synthetic speech. *Speech Communication*, 16, 369-390.

- Murray, I. R., Arnott, J. L., & Rohwer, E. A. (1996). Emotional stress in synthetic speech: Progress and future directions. *Speech Communication*, 20(1-2), 85-91.
doi:[https://doi.org/10.1016/S0167-6393\(96\)00046-5](https://doi.org/10.1016/S0167-6393(96)00046-5)
- Nakata, W., Koriyama, T., Takamichi, S., Saito, Y., Iijima, Y., Masumura, R., & Saruwatari, H. (2022). *Predicting VQVAE-based Character Acting Style from Quotation-Annotated Text for Audiobook Speech Synthesis*. Paper presented at INTERSPEECH 2022, Incheon, Korea.
- Narendra, N. P., Schuller, B., & Alku, P. (2021). The Detection of Parkinson's Disease From Speech Using Voice Source Information. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 1925-1936.
- Narendranath, M., Murthy, H. A., Rajendran, S., & Yegnanarayana, B. (1995). Transformation of formants for voice conversion using artificial neural networks. *Speech Communication*, 16(2), 207-216. doi:[https://doi.org/10.1016/0167-6393\(94\)00058-1](https://doi.org/10.1016/0167-6393(94)00058-1)
- Ní Chasaide, A. (1999). Irish. In *Handbook of the International Phonetic Association* (pp. 111-116): Cambridge University Press.
- 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 INTERSPEECH 2004, Jeju Island, Korea.
- Ní Chasaide, A., & Gobl, C. (2004b). *Voice quality and f_0 in prosody: towards a holistic account*. Paper presented at 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), 1-16.
- Ní Chasaide, A., Ní Chiaráin, N., Errity, R., Mroz, O., Ní hAonghusa, O., Ní Chasaide, S., . . . Barnes, E. (2023). *Mol an Óige: a phonological awareness and early literacy platform for Irish*. Paper presented at the 9th Workshop on Speech and Language Technology in Education (SLaTE), Dublin, Ireland.
- Ní Chasaide, A., Ní Chiaráin, N., Wendler, C., Berthelsen, H., Murphy, A., & Gobl, C. (2017). *The ABaIR initiative: Bringing Spoken Irish into the Digital Space*. Paper presented at INTERSPEECH 2017, Stockholm, Sweden.
- Ní Chasaide, A., Yanushevskaya, I., & Gobl, C. (2011). *Voice source dynamics in intonation*. Paper presented at the XVII International Congress of Phonetic Sciences (ICPhS XVII), Hong Kong, China.

- Ní Chasaide, A., Yanushevskaya, I., & Gobl, C. (2015). *Prosody of voice: declination, sentence mode and interaction with prominence*. Paper presented at the XVIIIth International Congress of Phonetic Sciences (ICPhS XVIII), Glasgow, UK.
- Ní Chasaide, A., Yanushevskaya, I., Kane, J., & Gobl, C. (2013). *The Voice Prominence Hypothesis: the Interplay of f_0 and Voice Source Features in Accentuation*. Paper presented at INTERSPEECH 2013, Lyon, France.
- Ní Chasaide, A., Barnes, E., Ní Chiaráin, N., McGuirk, R., Morrin, O., Nic Corcráin, M., & Cummins, J. (2022). *Challenges in assistive technology development for an endangered language: an Irish (Gaelic) perspective*. Paper presented at the Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022), Association for Computational Linguistics, Dublin, Ireland.
- Ní Chiaráin, N., & Ní Chasaide, A. (2015). *The TATL framework for CALL development*. Paper presented at the Critical CALL - Proceedings of the 2015 EUROCALL Conference, Padova, Italy.
- Ní Chiaráin, N., & Ní Chasaide, A. (2019). *An Scéalaí: autonomous learners harnessing speech and language technologies*. Paper presented at the 8th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2019), Graz, Austria.
- Ní Chiaráin, N., & Ní Chasaide, A. (2020). The Potential of Text-to-Speech Synthesis in Computer-Assisted Language Learning: A Minority Language Perspective. In A. Andujar (Ed.), *Recent Tools for Computer- and Mobile-Assisted Foreign Language Learning* (pp. 149-169). Hershey, PA: IGI Global.
- Noble, L., & Xu, Y. (2011). *Friendly speech and happy speech - are they the same?* Paper presented at the 17th International Congress of Phonetic Sciences (ICPhS XVII), Hong Kong.
- Nolan, F. (2006). Intonation. In B. Aarts & A. McMahon (Eds.), *The Handbook of English Linguistics*. Blackwell Publishing.
- Nygaard, L. C., & Lunders, E. (2002). Resolution of lexical ambiguity by emotional tone of voice. *Memory & cognition*, 30, 583-593. doi:10.3758/BF03194959
- Nygaard, L. C., & Queen, J. S. (2008). Communicating emotion: Linking affective prosody and word meaning. *Journal of Experimental Psychology: Human Perception and Performance*, 34(4), 1017-1030.
- O'Connor, J. D., & Arnold, G. F. (1973). *Intonation of Colloquial English* (2nd ed.). London: Longman.
- O'Reilly, M. (2014). *Sentence mode, alignment and focus in the intonation of Cois Fharraige, Inis Mor and Gaoth Dobhair Irish - A dual approach*. (PhD Thesis). Trinity College Dublin,
- O'Reilly, M., & Ní Chasaide, A. (2007). *Analysis of Intonation Contours in Portrayed Emotions Using the Fujisaki Model*. Paper presented at the Doctoral Consortium, The Second International Conference on Affective Computing and Intelligent Interaction, Lisbon, Portugal.
- O'Reilly, M., & Ní Chasaide, A. (2015). *Declination, peak height and pitch level in declaratives and questions of South Connaught Irish*. Paper presented at INTERSPEECH 2015, Dresden, Germany.

- O'Reilly, M., & Ní Chasaide, A. (2018). *IP length and peak (and valley) trends in neutral declaratives in Connaught and Ulster Irish – a comparison*. Paper presented at Speech Prosody 2018, Poznań, Poland.
- Ogden, R. (2001). Turn transition, creak and glottal stop in Finnish talk-in-interaction. *Journal of the International Phonetic Association*, 31(1), 139-152.
- Ohala, J. J. (1984). An Ethological Perspective on Common Cross-Language Utilization of F0 of Voice. *Phonetica: International Journal of Speech Science*, 41(1), 1-16. doi:10.1159/000261706
- Olivera, L. C. (1993). *Estimation of source parameters by frequency analysis*. Paper presented at the 3rd European Conference on Speech Communication and Technology (Eurospeech 1993), Berlin, Germany.
- Olivola, C. Y., Funk, F., & Todorov, A. (2014). Social attributions from faces bias human choices. *Trends in cognitive sciences*, 18(11), 566-570. doi:10.1016/j.tics.2014.09.007
- Ortony, A., & Turner, T. (1990). What's basic about basic emotions? *Psychological Review*, 97, 315-331.
- Paeschke, A. (2004). *Global Trend of Fundamental Frequency in Emotional Speech*. Paper presented at 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 (ICPhS XIV), San Francisco, USA.
- Patel, S., Scherer, K. R., Björkner, E., & Sundberg, J. (2011). Mapping emotions into acoustic space: The role of voice production. In (Vol. 87, pp. 93-98).
- Paulmann, S., Jessen, S., & Kotz, S. A. (2009). Investigating the multimodal nature of human communication: Insights from ERPs. *Journal of Psychophysiology*, 23(2), 63-76. doi:https://doi.org/10.1027/0269-8803.23.2.63
- Paulmann, S., & Kotz, S. A. (2008). An ERP investigation on the temporal dynamics of emotional prosody and emotional semantics in pseudo- and lexical-sentence context. *Brain and Language*, 105, 59-69.
- Paulmann, S., & Pell, M. D. (2010). Contextual influences of emotional speech prosody on face processing: How much is enough? *Cognitive, Affective, & Behavioral Neuroscience*, 10(2), 230-242. doi:10.3758/CABN.10.2.230
- Paulmann, S., & Pell, M. D. (2011). Is there an advantage for recognizing multi-modal emotional stimuli? *Motivation and Emotion*, 35(2), 192-201. doi:https://doi.org/10.1007/s11031-011-9206-0
- Pell, M. D. (2002). Evaluation of nonverbal emotion in face and voice: Some preliminary findings on a new battery of tests. *Brain and Cognition*, 48, 499-504.
- Pell, M. D., Jaywant, A., Monetta, L., & Kotz, S. A. (2011). Emotional speech processing: disentangling the effects of prosody and semantic cues. *Cognition & Emotion*, 25(5), 834-853. doi:10.1080/02699931.2010.516915

- Pell, M. D., & Kotz, S. A. (2011). On the Time Course of Vocal Emotion Recognition. *PLoS ONE*, 6(11), e27256. doi:10.1371/journal.pone.0027256
- Pell, M. D., Monetta, L., Paulmann, S., & Kotz, S. (2009). Recognizing Emotions in a Foreign Language. *Journal of Nonverbal Behavior*, 33, 107-120. doi:10.1007/s10919-008-0065-7
- Perrotin, O., & McLoughlin, I. (2019). *A spectral glottal flow model for source-filter separation of speech*. Paper presented at the ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, UK.
- Perrotin, O., & McLoughlin, I. V. (2017). On the Use of a Spectral Glottal Model for the Source-filter Separation of Speech. doi:arXiv:1712.08034
- Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation*. (PhD Thesis). Massachusetts Institute of Technology, Massachusetts
- Pisanski, K., Cartei, V., McGettigan, C., Raine, J., & Reby, D. (2016). Voice Modulation: A Window into the Origins of Human Vocal Control? *Trends in cognitive sciences*, 20(4), 304-318.
- Pitrelli, J. F., Bakis, R., Eide, E. M., Fernandez, R., Hamza, W., & Picheny, M. A. (2006). The IBM Expressive Text-to-Speech Synthesis System for American English. *IEEE Transactions on Audio Speech and Language Processing*, 14(4), 1099-1108. doi:10.1109/TASL.2006.876123
- 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, London: The Guilford Press.
- Plutchik, R. (1980). *Emotion : a Psychoevolutionary Synthesis* New York: Harper & Row.
- Plutchik, R. (1994). *The psychology and biology of emotion*. New York: Harper-Collins.
- Podesva, R. J. (2007). Phonation type as a stylistic variable: The use of falsetto in constructing a persona. *Journal of Sociolinguistics*, 11(4), 478-504.
- Podesva, R. J., & Callier, P. (2015). Voice Quality and Identity. *Annual Review of Applied Linguistics*, 35, 173-194. doi:10.1017/S0267190514000270
- Porter, H. C. (1963). Extraction of pitch from the trachea. *Air Force Cambridge Research Laboratories Research Note, AFCRL-63-24*.
- Posit Team. (2025). RStudio: Integrated Development Environment for R [Computer Software]. Boston, MA: Posit Software, PBC. <http://www.posit.co/>
- Prinz, J. J. (2004). *Gut Reactions: a Perceptual Theory of Emotion*. Oxford: Oxford University Press.
- Protopapas, A., & Lieberman, P. (1997). Fundamental frequency of phonation and perceived emotional stress. *Journal of The Acoustical Society of America*, 101(4), 2267-2277. doi:<https://doi.org/10.1121/1.418247>

- Qin, L., Ling, Z.-H., Wu, Y.-J., Zhang, B.-F., & Wang, R.-H. (2006). HMM-Based Emotional Speech Synthesis Using Average Emotion Model. In Q. Huo, B. Ma, E. S. Chng, & H. Li (Eds.), *Chinese Spoken Language Processing. ISCSLP 2006. Lecture Notes in Computer Science* (Vol. 4274, pp. 233-240). Berlin, Heidelberg: Springer.
- Qin, Y., Zhang, X., & Ying, H. (2010). *A HMM-based fuzzy affective model for emotional speech synthesis*. Paper presented at the 2010 2nd International Conference on Signal Processing Systems (ICSPS).
- Quené, H., Semin, G. R., & Foroni, F. (2012). Audible smiles and frowns affect speech comprehension. *Speech Communication, 54*(7), 917-922. doi:10.1016/j.specom.2012.03.004
- R Core Team. (2021). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>.
- Rachman, L., Liuni, M., Arias, P., Lind, A., Johansson, P., Hall, L., . . . Aucouturier, J.-J. (2018). DAVID: An open-source platform for real-time transformation of infra-segmental emotional cues in running speech. *Behavior Research Methods, 50*(1), 323-343. doi:<https://doi.org/10.3758/s13428-017-0873-y>
- Raitio, T., Li, J., & Seshadri, S. (2022, 23-27 May 2022). *Hierarchical Prosody Modeling and Control in Non-Autoregressive Parallel Neural TTS*. Paper presented at the ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore.
- Rank, E., & Pirker, H. (1998). *Generating emotional speech with a concatenative synthesizer*. Paper presented at the ICSLP.
- Raphael, L. J., Borden, G. J., & Harris, K. S. (2011). *Speech Science Primer: Physiology, Acoustics, and Perception of Speech* (6 ed.). Baltimore, MD: Lippincott Williams & Wilkins.
- Reddy, M. G., Harikrishna, D. M., Rao, S., & Manjunath, K. E. (2015, 2-3 Jan. 2015). *Telugu emotional story speech synthesis using SABLE markup language*. Paper presented at the 2015 International Conference on Signal Processing and Communication Engineering Systems.
- Redi, L., & Shattuck-Hufnagel, S. (2001). Variation in the realization of glottalization in normal speakers. *Journal of Phonetics, 29*(4), 407-429. doi:<https://doi.org/10.1006/jpho.2001.0145>
- Reetz, H., & Jongman, A. (2020). *Phonetics: Transcription, Production, Acoustics and Perception*. Hoboken: Wiley-Blackwell.
- Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2019). *FastSpeech: Fast, Robust and Controllable Text to Speech*. Paper presented at the Advances in Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.
- Richter, K. (2019). *English-Medium Instruction and Pronunciation: Exposure and Skills Development: Multilingual Matters*.
- Rigoulot, S., & Pell, M. D. (2014). Emotion in the voice influences the way we scan emotional faces. *Speech Communication, 65*, 36-49. doi:10.1016/j.specom.2014.05.006
- Rodgers, A. E. (2022). plugin_drawImportedParameters. Retrieved from https://github.com/AERodgers/plugin_drawImportedParameters

- Rosenberg, A. E. (1971). Effect of Glottal Pulse Shape on the Quality of Natural Vowels. *The Journal of the Acoustical Society of America*, 49(2B), 583-590. doi:10.1121/1.1912389
- Rossato, S., Audibert, N., & Aubergé, V. (2004). *Emotional voice measurement: a comparison of articulatory-EGG and acoustic-amplitude parameters*. Paper presented at Speech Prosody 2004, Nara, Japan.
- Rothenberg, M. (1972). The Glottal Volume Velocity Waveform during Loose and Tight Voiced Glottal Adjustments. In R. André & C. René (Eds.), *Proceedings of the seventh International Congress of Phonetic Sciences / Actes du Septième Congrès international des sciences phonétiques* (pp. 380-388). Berlin, Boston: De Gruyter Mouton.
- 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.
- Rubin, P., Baer, T., & Mermelstein, P. (1981). An articulatory synthesizer for perceptual research. *The Journal of the Acoustical Society of America*, 70(2), 321-328. doi:10.1121/1.386780
- 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 categorization of emotions. *Psychological Bulletin*, 110(3), 426-450.
- Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, 110(1), 145-172.
- Russell, J. A., Bachorowski, J.-A., & Fernandez-Dols, J.-M. (2003). Facial and Vocal Expressions of Emotion. *Annual Review of Psychology*, 54, 329-349. doi:10.1146/annurev.psych.54.101601.145102
- Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology*, 76(5), 805-819. doi:10.1037//0022-3514.76.5.805
- Ryan, C., Ní Chasaide, A., & Gobl, C. (2003). *Voice quality variation and the perception of affect: continuous or categorical?* Paper presented at the 15th International Congress of Phonetic Sciences (ICPhS XV), Barcelona, Spain.
- Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. L. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences of the United States of America*, 107, 2408-2412.
- Sauter, D. A., & Fischer, A. H. (2018). Can perceivers recognise emotions from spontaneous expressions? *Cognition & Emotion*, 32, 504-515. doi:https://doi.org/10.1080/02699931.2017.1320978
- Sauter, D. A., Panattoni, C., & Happé, F. (2013). Children's recognition of emotions from vocal cues *British Journal of Developmental Psychology*, 31, 97-113.
- Sauter, D. A., & Scott, S. K. (2007). More than one kind of happiness: Can we recognize vocal expressions of different positive states? *Motivation and Emotion*, 31, 192-199.

- Sawamura, K., Dang, J., Akagi, M., Erickson, D., Li, A., Sakuraba, K., . . . Hirose, K. (2007). *Common factors in emotion perception among different cultures*. Paper presented at the XVth International Congress of Phonetic Sciences (ICPhS XVI) , Saarbrücken, Germany.
- Scherer, K. R. (1974). Voice quality analysis of American and German speakers. *Journal of Psycholinguistic Research*, 3, 281-298. doi:<https://doi.org/10.1007/BF01069244>
- Scherer, K. R. (1981). Speech and Emotional States. In J. Darby (Ed.), *Speech evaluation in psychiatry* (pp. 189-220). New York: Grunne & Stratton.
- Scherer, K. R. (1982). Methods of research on vocal communication: paradigms and parameters. In K. R. Scherer & P. Ekman (Eds.), *Handbook of methods in nonverbal behavior research* (pp. 136-198). Cambridge, UK: Cambridge University Press.
- Scherer, K. R. (1986). Vocal affect expression: a review and a model for future research. *Psychological Bulletin*, 99(2), 143-165.
- Scherer, K. R. (1988). On the Symbolic Functions of Vocal Affect Expression. *Journal of Language and Social Psychology*, 7(2), 79-100. doi:10.1177/0261927x8800700201
- 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. (1996). *Adding the affective dimension: A new look in speech analysis and synthesis*. Paper presented at the 4th International Conference of Spoken Language Processing, Philadelphia, USA.
- Scherer, K. R. (1999). Appraisal theory. In T. Dalgleish & M. Power (Eds.), *Handbook of Cognition and Emotion* (pp. 637-663). New York: John Wilay & Sons.
- 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* (3 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. (2001). Appraisal considered as a process of multilevel sequential checking. In K. R. Scherer, A. Schorr, & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, methods, research* (pp. 92-120). Oxford: Oxford University Press.
- Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1), 227-256. doi:[https://doi.org/10.1016/S0167-6393\(02\)00084-5](https://doi.org/10.1016/S0167-6393(02)00084-5)
- 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. (2013). Vocal markers of emotion: Comparing induction and acting elicitation. *Computer Speech & Language*, 27, 40-58.

- Scherer, K. R. (2021). Comment: Advances in studying the vocal expression of emotion: Current contributions and further options. *Emotion Review*, 13(1), 57-59.
- 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., 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).
- Scherer, K. R., Koivumaki, J., & Rosenthal, R. (1972). Minimal cues in the vocal communication of affect: Judging emotions from content-masked speech. *Journal of Psycholinguistic Research*, 1(3), 269-285. doi:<https://doi.org/10.1007/BF01074443>
- Scherer, K. R., Ladd, D. R., & Silverman, K. E. A. (1984). Vocal Cues to Speaker Affect: Testing two models. *Journal of The Acoustical Society of America*, 76. doi:10.1121/1.391450
- Scherer, K. R., & Oshinsky, J. S. (1977). Cue Utilization in Emotion Attribution from Auditory Stimuli. *Motivation and Emotion*, 1(4), 331-346.
- Scherer, K. R., Rosenthal, R., & Koivumaki, J. (1972). Mediating interpersonal expectancies via vocal cues: Differential speech intensity as a means of social influence. *European Journal of Social Psychology*, 2(2), 163-175. doi:10.1002/ejsp.2420020205
- Scherer, K. R., & Scherer, U. (1981). Speech behavior and personality. In J. Darby (Ed.), *Speech evaluation in psychiatry* (pp. 115-135). New York: Grunne & Stratton.
- Scherer, K. R., & Scherer, U. (2011). Assessing the Ability to Recognize Facial and Vocal Expressions of Emotion: Construction and Validation of the Emotion Recognition Index. *Journal of Nonverbal Behavior*, 35, 305-326.
- 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, 310-328. doi:<https://doi.org/10.1037/0022-3514.66.2.310>
- Schlosberg, H. (1954). Three dimensions of emotion. *Psychological Review*, 61(2), 81-88. doi:<https://doi.org/10.1037/h0054570>
- 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: a review*. Paper presented at INTERSPEECH 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). Kloster Irsee, Germany: Springer.

- Schröder, M. (2004b). *Speech and Emotion Research: an Overview of Frameworks and a Dimensional Approach to Emotional Speech Synthesis*. (PhD Thesis). Saarland University, Research Report of the Institute of Phonetics.
- Schröder, M. (2009). Expressive Speech Synthesis: Past, Present, and Possible Futures. In J. Tao & T. Tan (Eds.), *Affective Information Processing* (pp. 111-126). London: Springer London.
- 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 7th European Conference on Speech Communication and Technology - EUROSPEECH 2001, Aalborg, Denmark.
- Shariff, A., & Tracy, J. (2011). What are emotion expressions for? *Current Directions in Psychological Science*, 20, 395-399.
- Shaver, P., 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. doi:<https://doi.org/10.1037/0022-3514.52.6.1061>
- Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., . . . Wu, Y. (2018). Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4779-4783.
- Shen, K., Ju, Z., Tan, X., Liu, Y., Leng, Y., He, L., . . . Bian, J. (2023). NaturalSpeech 2: Latent Diffusion Models are Natural and Zero-Shot Speech and Singing Synthesizers. doi:<https://doi.org/10.48550/arXiv.2304.09116>
- Shikano, K., Nakamura, S., & Abe, M. (1991). *Speaker adaptation and voice conversion by codebook mapping*. Paper presented at the 1991 IEEE International Symposium on Circuits and Systems, Singapore.
- Shue, Y.-L., & Alwan, A. (2010). *A new voice source model based on high-speed imaging and its application to voice source estimation*. Paper presented at the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing.
- Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., . . . Hirschberg, J. (1992). *ToBI: A standard for labeling English prosody*. Paper presented at the Second International Conference on Spoken Language Processing (ICSLP), Banff, Canada.
- Simon-Thomas, E. R., Keltner, D. J., Sauter, D., Sinicropi-Yao, L., & Abramson, A. (2009). The Voice Conveys Specific Emotions: Evidence From Vocal Burst Displays. *Emotion*, 9(6), 838-846.
- Sisman, B., Yamagishi, J., King, S., & Li, H. (2020). An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29, 132–157. doi:10.1109/taslp.2020.3038524
- Sjölander, K., & Beskow, J. (2000). *Wavesurfer - an open source speech tool*. Paper presented at the 6th International Conference on Spoken Language Processing (ICSLP 2000).
- Skerry-Ryan, R. J., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., . . . Saurous, R. A. (2018). *Towards end-to-end prosody transfer for expressive speech synthesis with tacotron*. Paper presented at the 35th International Conference on Machine Learning (PMLR 80, 2018), Stockholm, Sweden.

- Sluijter, A. M. C., Shattuck-Hufnagel, S., Stevens, K. N., & van Heuven, V. J. (1995). *Supralaryngeal resonance and glottal pulse shape as correlates 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). *Acoustic correlates of linguistic stress and accent in Dutch and American English*. Paper presented at the Fourth International Conference on Spoken Language Processing (ICSLP '96), Philadelphia, PA, USA
- Sluijter, A. M. C., van Heuven, V. J., & Pacilly, J. J. (1997). Spectral balance as a cue in the perception of linguistic stress. *Journal of The Acoustical Society of America*, 101(1), 503-513. doi:10.1121/1.417994
- 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.
- Sotelo, J., Mehri, S., Kumar, M., Santos, J. F., Kastner, K., Courville, A., & Bengio, Y. (2017). *Char2Wav: End-to-End Speech Synthesis*. *International Conference on Learning Representations*. Paper presented at the 5th International Conference on Learning Representations (ICLR 2017), Toulon, France.
- Starkweather, J. A. (1956). Content-free speech as a source of information about the speaker. *The Journal of Abnormal and Social Psychology*, 52(3), 394-402. doi:https://doi.org/10.1037/h0041133
- Stevens, K. N. (1999). *Acoustic Phonetics*: The MIT Press.
- Strawbridge, T., Soneson, D., & Griffith, C. (2019). Lasting effects of pre-university language exposure on undergraduate proficiency. 52(4), 776-797.
- Strik, H., & Boves, L. (1994). *Automatic estimation of voice source parameters*. Paper presented at the 3rd International Conference on Spoken Language Processing (ICSLP 1994), Yokohama, Japan.
- Stuart-Smith, J. (1999). Glasgow: Accent and voice quality. In P. Foulkes & G. J. Docherty (Eds.), *Urban voices: Accent Studies in the British Isles* (Vol. Leeds, UK, pp. 201-222): Arnold.
- Styger, T., & Keller, E. (1994). Formant Synthesis. In E. Keller (Ed.), *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art, and Future Challenges* (pp. 109-128). Chichester: John Wiley
- Stylianou, Y. (2001). Applying the harmonic plus noise model inconcatenative speech synthesis. *IEEE Transactions on Speech and Audio Processing*, 9(1), 21-29.
- Stylianou, Y. (2009). *Voice Transformation: A survey*. Paper presented at the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing.
- Suzié, S., Nosek, T., Sečujski, M., Pekar, D., & Delié, V. (2019, 26-27 Nov. 2019). *DNN Based Expressive Text-to-Speech with Limited Training Data*. Paper presented at the 2019 27th Telecommunications Forum (TELFOR), Belgrade, Serbia.

- Swerts, M., & Veldhuis, R. (2001). The effect of speech melody on voice quality. *Speech Communication, 33*(4), 297-303. doi:[https://doi.org/10.1016/S0167-6393\(00\)00061-3](https://doi.org/10.1016/S0167-6393(00)00061-3)
- Szakay, A. (2012). Voice quality as a marker of ethnicity in New Zealand: From acoustics to perception. *Journal of Sociolinguistics, 16*(3), 382-397.
- Tanaka, A., Koizumi, A., Imai, H., Hiramatsu, S., Hiramoto, E., & de Gelder, B. (2010). I Feel Your Voice: Cultural Differences in the Multisensory Perception of Emotion. *Psychological Science, 21*, 1259-1262. doi:10.1177/0956797610380698
- Tao, J., Kang, Y., & Li, A. (2006). Prosody conversion from neutral speech to emotional speech. *IEEE Transactions on Audio Speech and Language Processing, 14*(4), 1145-1154.
- Terken, J. (1991). Fundamental frequency and perceived prominence of accented syllables. *Journal of The Acoustical Society of America, 89*(4), 1768-1776.
- Terken, J. (1994). Fundamental frequency and perceived prominence of accented syllables. II. Nonfinal accents. *Journal of The Acoustical Society of America, 95*(6), 3662-3665.
- Teshigawara, M. (2004). *Random splicing: a method of investigating the effects of voice quality on impression formation*. Paper presented at Speech Prosody 2004, Nara, Japan.
- The Mathworks Inc. (2022). MATLAB (Version 9.12.0.2170939 (R2022a)) [Computer Software]. Natick, Massachussets: The MathWorks Inc. <https://www.mathworks.com>
- Thompson, A. E., & Voyer, D. (2014). Sex differences in the ability to recognise non-verbal displays of emotion: A meta-analysis. *Cognition and Emotion, 28*(7), 1164-1195. doi:10.1080/02699931.2013.875889
- Thompson, W. F., & Balkwill, L.-L. (2006). Decoding speech prosody in five languages. *Semiotica, 158*(1/4), 407-424. doi:10.1515/SEM.2006.017
- Thurgood, E. (2004). Phonation Types in Javanese. *Oceanic Linguistics, 43*(2), 277-295.
- Tits, N. (2019). A Methodology for Controlling the Emotional Expressiveness in Synthetic Speech - a Deep Learning approach. *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, 1-5.
- Tits, N., El Haddad, K., & Dutoit, T. (2019). *Exploring Transfer Learning for Low Resource Emotional TTS*. Paper presented at the IntelliSys- Intelligent Systems Conference.
- Titze, I. R., Svec, J. G., & Popolo, P. S. (2003). Vocal dose measures: quantifying accumulated vibration exposure in vocal fold tissues. *J Speech Lang Hear Res, 46*(4), 919-932. doi:10.1044/1092-4388(2003/072)
- Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., & Oura, K. (2013). Speech Synthesis Based on Hidden Markov Models. *Proceedings of the IEEE, 101*(5), 1234-1252. doi:10.1109/JPROC.2013.2251852
- Tomkins, S. S. (1995). *Exploring Affect: the Selected Writings of Sylvan S. Tomkins*. Cambridge: Cambridge University Press.

- Torre, I. (2014). *Production and perception of smiling voice*. Paper presented at the First Postgraduate and Academic Researchers in Linguistics at York (PARLAY 2013) Conference, York, United Kingdom.
- Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., . . . Nelson, C. (2009). The NimStim set of facial expressions: judgments from untrained research participants. *Psychiatry Res*, 168(3), 242-249. doi:10.1016/j.psychres.2008.05.006
- Triandafilidi, I. I., Tatarnikova, T. M., & Poponin, A. S. (2022). *Speech Synthesis System for People with Disabilities*. Paper presented at the Wave Electronics and its Application in Information and Telecommunication Systems (WECONF), Saint Petersburg, Russian Federation.
- Triantafyllopoulos, A., Schuller, B. W., İymen, G., Sezgin, M., He, X., Yang, Z., . . . Tao, J. (2023). An Overview of Affective Speech Synthesis and Conversion in the Deep Learning Era. *Proceedings of the IEEE*, 111(10), 1355-1381. doi:10.1109/JPROC.2023.3250266
- Trudgill, P. W. (1974). *The Social Differentiation of English in Norwich*. Cambridge, UK: Cambridge University Press.
- Tsai, J. L., Knutson, B., & Fung, H. H. (2006). Cultural variation in affect valuation. *Journal of Personality and Social Psychology*, 90(2), 288-307. doi:https://doi.org/10.1037/0022-3514.90.2.288
- Tsai, J. L., Louie, J. Y., Chen, E. E., & Uchida, Y. (2007). Learning What Feelings to Desire: Socialization of Ideal Affect Through Children's Storybooks. *Personality and Social Psychology Bulletin*, 33(1), 17-30. doi:https://doi.org/10.1177/0146167206292749
- Türk, O., & Schröder, M. (2010). Evaluation of Expressive Speech Synthesis With Voice Conversion and Copy Resynthesis Techniques. *IEEE Transactions on Audio Speech and Language Processing*, 18(5), 965-973.
- Uldall, E. (1964). Dimensions of meaning in intonation. In D. Abercrombie, D. B. Fry, P. A. D. MacCarthy, N. 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). London, England: Longmans.
- Um, S.-Y., Oh, S., Byun, K., Jang, I., Ahn, C., & Kang, H.-G. (2020). *Emotional Speech Synthesis with Rich and Granularized Control*. Paper presented at the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain.
- UNESCO. (2010). *Atlas of the world's languages in danger* (3 ed.). Paris: UNESCO Publishing.
- Vainio, M., Airas, M., Järvikivi, J., & Alku, P. (2010). *Laryngeal voice quality in the expression of focus*. Paper presented at INTERSPEECH 2010, Makuhari, Chiba, Japan.
- Vainio, M., & Järvikivi, J. (2006). Tonal features, intensity, and word order in the perception of prominence. *Journal of Phonetics*, 34(3), 319-342. doi:10.1016/j.wocn.2005.06.004

- Valle, R., Li, J., Prenger, R., & Catanzaro, B. (2020). *Mellotron: Multispeaker Expressive Voice Synthesis by Conditioning on Rhythm, Pitch and Global Style Tokens*. Paper presented at the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain.
- van Bezooijen, R. (1984). *Characteristics and recognizability of vocal expression of emotion*. Dordrecht, the Netherlands: Foris.
- van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., . . . Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *Proceedings of the 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, 125. doi:10.1109/ICASSP.2009.4960364
- van Rijn, P., & Larrouy-Maestri, P. (2023). Modelling individual and cross-cultural variation in the mapping of emotions to speech prosody. *Nature Human Behaviour*, 7(3), 386-396. doi:10.1038/s41562-022-01505-5
- Veldhuis, R. (1998). A computationally efficient alternative for the Liljencrants-Fant model and its perceptual evaluation. *J Acoust Soc Am*, 103(1), 566-571. doi:10.1121/1.421103
- Velten, E., Jr. (1968). A laboratory task for induction of mood states. *Behaviour Research and Therapy*, 6(4), 473-482. doi:https://doi.org/10.1016/0005-7967(68)90028-4
- Vioni, A., Christidou, M., Ellinas, N., Vamvoukakis, G., Kakoulidis, P., Kim, T., . . . Tsiakoulis, P. (2021). *Prosodic Clustering for Phoneme-Level Prosody Control in End-to-End Speech Synthesis*. Paper presented at the ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada.
- Vroomen, J., Collier, R., & Mozziconacci, S. (1993). *Duration and intonation in emotional speech*. Paper presented at the Eurospeech 1993, Berlin, Germany.
- Waaramaa, T. (2014). Perception of emotional nonsense sentences in China, Egypt, Estonia, Finland, Russia, Sweden, and the USA. *Logopedics, phoniatrics, vocology*, 40, 1-7. doi:10.3109/14015439.2014.915982
- Waaramaa, T., Alku, P., & Laukkanen, A.-M. (2006). The role of F3 in the vocal expression of emotions. *Logopedics, phoniatrics, vocology*, 31(4), 153-156. doi:https://doi.org/10.1080/14015430500456739
- Waaramaa, T., & Leisiö, T. (2013). Perception of emotionally loaded vocal expressions and its connection to responses to music. A cross-cultural investigation: Estonia, Finland, Sweden, Russia, and the USA. *Frontiers in Psychology*, 4, 344. doi:10.3389/fpsyg.2013.00344
- Wagner, H. L. (1993). On measuring performance in category judgment studies of nonverbal behavior. *Journal of Nonverbal Behavior*, 17, 3-28. doi:https://doi.org/10.1007/BF00987006
- Wambacq, I. J. A., & Jerger, J. F. (2004). Processing of affective prosody and lexical-semantics in spoken utterances as differentiated by event-related potentials. *Cognitive Brain Research*, 20(3), 427-437. doi:10.1016/j.cogbrainres.2004.03.015
- Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., . . . Wei, F. (2023). Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. doi: https://doi.org/10.48550/arXiv.2301.02111

- Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., . . . Saurous, R. A. (2017). *Tacotron: Towards End-to-End Speech Synthesis*. Paper presented at INTERSPEECH 2017, Stockholm, Sweden.
- Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R. J., Battenberg, E., Shor, J., . . . Saurous, R. A. (2018). *Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis*. Paper presented at the 35th International Conference on Machine Learning (PMLR 80, 2018), Stockholm, Sweden.
- Wang, Z., & Gobl, C. (2022). *Contribution of the glottal flow residual in affect-related voice transformation*. Paper presented at INTERSPEECH 2022, Incheon, South Korea.
- Wang, Z., & Gobl, C. (2023a). *Perceptual effects of aliasing distortion in glottal flow modelling*. Paper presented at the 20th International Congress of Phonetic Sciences (ICPhS XX), Prague, Czech Republic.
- Wang, Z., & Gobl, C. (2023b). *A System for Generating Voice Source Signals that Implements the Transformed LF-model Parameter Control*. Paper presented at INTERSPEECH 2023, Dublin, Ireland.
- Wells, J. C. (2006). *English Intonation: An Introduction*. Cambridge: Cambridge University Press.
- 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.
- Widen, S. C., & Russell, J. A. (2008). Children acquire emotion categories gradually. *Cognitive Development*, 23(2), 291-312. doi:<https://doi.org/10.1016/j.cogdev.2008.01.002>
- 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. actor emotional speech*. Paper presented at INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA.
- Włodarczak, M., Ludusan, B., Sundberg, J., & Heldner, M. (2025). Classification of Voice Quality Using Neck-Surface Acceleration: Comparison With Glottal Flow and Radiated Sound. *Journal of Voice*, 39(1), 10-24. doi:<https://doi.org/10.1016/j.jvoice.2022.06.034>
- Wong, D. Y., Markel, J. D., & Gray, A. H. (1979). Least squares glottal inverse filtering from the acoustic speech waveform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(4), 350-355. doi:10.1109/TASSP.1979.1163260
- Wu, P., Ling, Z., Liu, L., Jiang, Y., Wu, H., & Dai, L. (2019, 18-21 Nov. 2019). *End-to-End Emotional Speech Synthesis Using Style Tokens and Semi-Supervised Training*. Paper presented at the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China.
- Wu, Y.-C., Hayashi, T., Tobing, P. L., Kobayashi, K., & Toda, T. (2019). *Quasi-Periodic WaveNet Vocoder: A Pitch Dependent Dilated Convolution Model for Parametric Speech Generation*. Paper presented at INTERSPEECH 2019, Graz, Austria.

- Wu, Z., Ning, Y., Zang, X., Jia, J., Meng, F., Meng, H. M., & Cai, L. (2014). Generating emphatic speech with hidden Markov model for expressive speech synthesis. *Multimedia Tools and Applications*, 74, 9909-9925.
- Xue, L., Zhu, X., An, X., & Xie, L. (2018). *A Comparison of Expressive Speech Synthesis Approaches based on Neural Network*. Paper presented at the ASMMC-MMAC'18: the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data, Seoul, Republic of Korea.
- Yamagishi, J., Masuko, T., & Kobayashi, T. (2004). *HMM-Based Expressive Speech Synthesis - Towards TTS with Arbitrary Speaking Styles and Emotions*. Paper presented at the Special Workshop in Maui (SWIM), Maui, Hawaii.
- Yamagishi, J., Nose, T., Zen, H., Ling, Z.-H., Toda, T., Tokuda, K., . . . Renals, S. (2009). Robust Speaker-Adaptive HMM-based Text-to-Speech Synthesis. *IEEE Transactions on Audio, Speech and Language Processing*, 17(6), 1208-1230. doi:10.1109/TASL.2009.2016394
- Yamamoto, R., Song, E., & Kim, J.-M. (2020, 4-8 May 2020). *Parallel Wavegan: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram*. Paper presented at the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020), Barcelona, Spain.
- Yanushevskaya, I. (2010). *Vocal correlates of affective states*. (PhD Thesis). Trinity College Dublin, Dublin.
- Yanushevskaya, I., Gobl, C., Kane, J., & Ní Chasaide, A. (2010). *An exploration of voice source correlates of focus*. Paper presented at INTERSPEECH 2010, Makuhari, Chiba, Japan.
- Yanushevskaya, I., Gobl, C., & Ní Chasaide, A. (2009). *Voice parameter dynamics in portrayed emotions*. Paper presented at the 6th International Workshop on Models and Analysis of Vocal Emissions for Biometrical Applications (MAVEBA 2009), Florence, Italy.
- Yanushevskaya, I., Gobl, C., & Ní Chasaide, A. (2018). Cross-language differences in how voice quality and f_0 contours map to affect. *Journal of The Acoustical Society of America*, 144(5), 2730-2750. doi:10.1121/1.5066448
- Yanushevskaya, I., Murphy, A., Gobl, C., & Ní Chasaide, A. (2016). *Perceptual salience of voice source parameters in signaling focal prominence*. Paper presented at INTERSPEECH 2016, San Francisco, USA.
- Yanushevskaya, I., Ní Chasaide, A., & Gobl, C. (2011). *Universal and language-specific perception of affect from voice*. Paper presented at the The 17th International Congress of Phonetic Sciences (ICPhS XVII), Hong Kong, China.
- Yanushevskaya, I., Ní Chasaide, A., & Gobl, C. (2017). *Cross-speaker Variation in Voice Source Correlates of Focus and Deaccentuation*. Paper presented at INTERSPEECH 2017, Stockholm, Sweden.
- Yilmazyildiz, S., Mattheyses, W., Patsis, Y., & Verhelst, W. (2006). *Expressive Speech Recognition and Synthesis as Enabling Technologies for Affective Robot-Child Communication*. Paper presented at the Advances in Multimedia Information Processing - PCM 2006.

- Yoon, H.-W., Kwon, O., Lee, H., Yamamoto, R., Song, E., Kim, J.-M., & Hwang, M.-J. (2022). *Language Model-Based Emotion Prediction Methods for Emotional Speech Synthesis Systems*. Paper presented at INTERSPEECH 2022, Incheon, Korea.
- Young, J. C., & Shishido, M. (2022). *Evaluating WaveNet Synthetic Speech for English as Second Language Listening Activities*. Paper presented at the 2022 Joint 12th International Conference on Soft Computing and Intelligent Systems and 23rd International Symposium on Advanced Intelligent Systems (SCIS&ISIS), Ise, Japan.
- Yuki, M., Maddux, W., & Masuda, T. (2007). Are the windows to the soul the same in the East and West? Cultural differences in using the eyes and mouth as cues to recognize emotions in Japan and the United States. *Journal of Experimental Social Psychology*, 43(2), 303-311. doi:10.1016/j.jesp.2006.02.004
- 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. C. Paiva, R. Prada, & R. W. Picard (Eds.), *Affective Computing and Intelligent Interaction. ACII 2007. Lecture Notes in Computer Science* (Vol. 4738, pp. 464-475). Berlin, Heidelberg: Springer Verlag.
- Zarins, U. (2017). *Anatomy of Facial Expression* (1 ed.). Digital Version: Anatomy Next, Inc.
- Zen, H., Tokuda, K., & Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication*, 51(11), 1039-1064. doi:10.1016/j.specom.2009.04.004
- Zhang, S. (2008, 2008//). *Emotion Recognition in Chinese Natural Speech by Combining Prosody and Voice Quality Features*. Paper presented at the Advances in Neural Networks - ISNN 2008, Berlin, Heidelberg.
- Zhang, S., Xu, Y., Jia, J., & Cai, L. (2008, 16-19 Dec. 2008). *Analysis and Modeling of Affective Audio Visual Speech Based on PAD Emotion Space*. Paper presented at the 2008 6th International Symposium on Chinese Spoken Language Processing.
- Zhang, Z. (2016). Cause-effect relationship between vocal fold physiology and voice production in a three-dimensional phonation model. *J Acoust Soc Am*, 139(4), 1493. doi:10.1121/1.4944754
- Zhang, Z., Xiao, Z., & Huang, M. (2022). Determination of the vocal tract model order in iterative adaptive inverse filtering. *Cognitive Computation and Systems*, 4(2), 138-146. doi:https://doi.org/10.1049/ccs2.12054
- Zhou, K., Sisman, B., Schuller, B. W., & Li, H. (2023). Emotion Intensity and its Control for Emotional Voice Conversion. *IEEE Transactions on Affective Computing*, 14(1), 31-48. doi:10.1109/TAFFC.2022.3175578
- Zuckermann, G. a., Vigfusson, S., Rayner, M., Ní Chiaráin, N., Ivanova, N., Habibi, H., & Bédi, B. (2021). LARA in the Service of Revivalistics and Documentary Linguistics: Community Engagement and Endangered Languages. *Proceedings of the Workshop on Computational Methods for Endangered Languages*, 1. doi:10.33011/computel.v1i.953

APPENDICES

Appendix A

Intonation Annotation Tables – All Recorded Sentences

All audio files analyzed in the tables below can be found and downloaded from:

<https://www.dropbox.com/sh/qqqd2w1pmccm7ji/AABg3OTT5bJ7XmyFpzcBxf-2a?dl=0>

Table A1 - Intonational Annotation of Sentence 1 (Beidh sé a' báistí ar an lá sin). The *Intonation* column provides the IViE labels at the phonological tier, followed by its respective pitch accent. The *Stress Pattern* column specifies the pitch accents of each recorded utterance (not necessarily reflecting the typical dialect intonation pattern).

SENTENCE 01 – Beidh sé a' báistí ar an lá sin			
Affect	Repetition	Intonation	Stress Pattern
Neutral	Rep 1	!H*L [báis]	%<báis>%
	Rep 2	!H*L [báis]	%<báis>%
	Rep 3	!H*L [báis]	%<báis>%
	Rep 4	!H*L [báis]	%<báis>%
	Rep 5	!H*L [báis]	%<báis>%
Angry	Rep 1	H*L [tí]	%<báis><tí>%
	Rep 2	H*L [tí]	%<báis><tí>%
	Rep 3	H*L [báis], H*L [lá]	%<báis><lá>% *matches neutral
	Rep 4	H*L [báis], H*L [lá]	%<báis><lá>% *matches neutral
	Rep 5	H*L [tí], H*L [lá]	%<tí><lá>%

Table A2 - Intonational Annotation of Sentence 2 (Gheobhmid bád ar an lá sin). The *Intonation* column provides the IViE labels at the phonological tier, followed by its respective pitch accent. The *Stress Pattern* column specifies the pitch accents of each recorded utterance (not necessarily reflecting the typical dialect intonation pattern).

SENTENCE 02 – Gheobhmid bád ar an lá sin			
Affect	Repetition	Intonation	Stress Pattern
Neutral	Rep 1	H* [gheo], !H* [bád], H*L [lá]	%<gheo><bád><lá>%
	Rep 2	H* [gheo], !H* [bád], H*L [lá]	%<gheo><bád><lá>%
	Rep 3	H* [gheo], !H*L [bád], H*L [lá]	%<gheo><bád><lá>%
	Rep 4	H* [gheo], !H*L [bád], H*L [lá]	%<gheo><bád><lá>%
	Rep 5	H* [gheo], !H*L [bád], H*L [lá]	%<gheo><bád><lá>%

Angry	Rep 1	H* [gheo], !H*L [bád], H*L [lá]	%<gheo><bád><lá>%
	Rep 2	H* [gheo], !H*L [bád], H*L [lá]	%<gheo><bád><lá>%
	Rep 3	H* [gheo], !H*L [bád], H*L [lá]	%<gheo><bád><lá>%
	Rep 4	H* [gheo], !H*L [bád], H*L [lá]	%<gheo><bád><lá>%
Bored	Rep 1	H*L [bád]	%<bád><lá>%
	Rep 2	H*L [bád]	%<bád><lá>%
	Rep 3	H*L [bád]	%<bád><lá>%
	Rep 4	H*L [bád], !H* [lá]	%<bád><lá>%
	Rep 5	H*L [bád], !H* [lá]	%<bád><lá>%
Happy	Rep 1	H*L [bád], !H*L [lá]	%<bád><lá>%
	Rep 2	H*L [bád], !H*L [lá]	%<bád><lá>%
	Rep 3	H*L [bád], !H*L [lá]	%<bád><lá>%
	Rep 4	H*L [bád], !H*L [lá]	%<bád><lá>%
	Rep 5	H*L [bád], !H*L [lá]	%<bád><lá>%
Interested	Rep 1	H*L [gheo], !H*L [bád], H*L [lá]	%<gheo><bád><lá>%
	Rep 2	H*L [gheo], !H*L [bád], H*L [lá]	%<gheo><bád><lá>%
	Rep 3	H*L [gheo], !H*L [bád], H*L [lá]	%<gheo><bád><lá>%
	Rep 4	H*L [gheo], !H*L [bád], H*L [lá]	%<gheo><bád><lá>%
	Rep 5	H*L [gheo], !H*L [bád], H*L [lá]	%<gheo><bád><lá>%
	Rep 6	H*L [gheo], !H*L [bád], H*L [lá]	%<gheo><bád><lá>%
Relaxed	Rep 1	H*L [gheo], L* [bád], H*L [lá]	%<gheo><bád><lá>%
	Rep 2	H*L [gheo], !H*L [bád], H*L [lá]	%<gheo><bád><lá>%
	Rep 3	H*L [gheo], !H*L [bád], H*L [lá]	%<gheo><bád><lá>%
	Rep 4	H*L [gheo], !H*L [bád], H*L [lá]	%<gheo><bád><lá>%
	Rep 5	H*L [gheo], L* [bád], H*L [lá]	%<gheo><bád><lá>%
	Rep 6	H*L [gheo], !H*L [bád], H*L [lá]	%<gheo><bád><lá>%
Sad	Rep 1	H*L [bád], H*L [lá]	%<bád><lá>%
	Rep 2	H*L [bád], H*L [lá]	%<bád><lá>%
	Rep 3*	H*L [bád], H*L [lá]	%<bád><lá>%
	Rep 4	H*L [bád], !H*L [lá]	%<bád><lá>%
	Rep 5	H*L [bád], H*L [lá]	%<bád><lá>%
	Rep 6	H*L [bád], !H*L [lá]	%<bád><lá>%

*Repetition 3 included variation added by the speaker.

Table A3 - Intonational Annotation of Sentence 3 (Beidh sé a' báistí amárach). The *Intonation* column provides the IVIE labels at the phonological tier, followed by its respective pitch accent. The *Stress Pattern* column specifies the pitch accents of each recorded utterance (not necessarily reflecting the typical dialect intonation pattern).

SENTENCE 03 – Beidh sé a' báistí amárach			
Affect	Repetition	Intonation	Stress Pattern
Neutral	Rep 1	H*L [báis], H*L [má]	%<báis><má>%
	Rep 2	H*L [báis], H*L [má]	%<báis><má>%
	Rep 3	H*L [báis], H*L [má]	%<báis><má>%
	Rep 4	H*L [báis], H*L [má]	%<báis><má>%
	Rep 5	H*L [báis], H*L [má]	%<báis><má>%
	Rep 6	H*L [báis], H*L [má]	%<báis><má>%
Angry	Rep 1	H* [beidh], H*L [báis], H*L [má]	%<beidh><báis><má>%
	Rep 2	H* [beidh], H*L [báis], H*L [má]	%<beidh><báis><má>%
	Rep 3	H*L [báis], H*L [má]	%<báis><má>%
	Rep 4	H*L [báis], H*L [má]	%<báis><má>%
	Rep 5	H*L [báis], H*% [má]	%<báis><má>%

Table A4 - Intonational Annotation of Sentence 4 (Beidh Barra ar an bád leo). The *Intonation* column provides the IVIE labels at the phonological tier, followed by its respective pitch accent. The *Stress Pattern* column specifies the pitch accents of each recorded utterance (not necessarily reflecting the typical dialect intonation pattern).

SENTENCE 04 – Beidh Barra ar an bád leo			
Affect	Repetition	Intonation	Stress Pattern
Neutral	Rep 1	H*L [Ba], H*L [bád]	%<Ba><bád>%
	Rep 2	H*L [Ba], H*L [bád]	%<Ba><bád>%
	Rep 3	H*L [Ba], H*L [bád]	%<Ba><bád>%
	Rep 4	H*L [Ba], H*L [bád]	%<Ba><bád>%
	Rep 5	H*L [Ba], H*L [bád]	%<Ba><bád>%
Angry	Rep 1	H*L [Ba], H*L [bád]	%<Ba><bád>%
	Rep 2	H*L [Ba], H*L [bád]	%<Ba><bád>%
	Rep 3	H*L [Ba], H*L [bád]	%<Ba><bád>%
	Rep 4	H*L [Ba], H*L [bád]	%<Ba><bád>%
	Rep 5	H*L [Ba], H*L [bád]	%<Ba><bád>%

Table A5 –Intonational Annotation of Sentence 5 (Beidh bád ann amárach). The *Intonation* column provides the IViE labels at the phonological tier, followed by its respective pitch accent. The *Stress Pattern* column specifies the pitch accents of each recorded utterance (not necessarily reflecting the typical dialect intonation pattern).

SENTENCE 05 – Beidh bád ann amárach			
Affect	Repetition	Intonation	Stress Pattern
Neutral	Rep 1	H* [beidh], !H* [bád], !H*L [má]	%<beidh><bád><má>%
	Rep 2	H* [beidh], !H* [bád], !H*L [má]	%<beidh><bád><má>%
	Rep 3	H* [beidh], !H* [bád], !H*L [má]	%<beidh><bád><má>%
	Rep 4	H* [beidh], !H* [bád], !H*L [má]	%<beidh><bád><má>%
	Rep 5	H* [beidh], !H* [bád], !H*L [má]	%<beidh><bád><má>%
	Rep 6	H* [beidh], !H* [bád], !H*L [má]	%<beidh><bád><má>%
Angry	Rep 1	H*L [bád], H*L [má]	%<bád><má>%
	Rep 2	H*L [bád], H*L [má]	%<bád><má>%
	Rep 3	H*L [bád], H*L [má]	%<bád><má>%
	Rep 4	H*L [bád], H*L [má]	%<bád><má>%
	Rep 5	H*L [bád], H*L [má]	%<bád><má>%

Table A6 - Intonational Annotation of Sentence 6 (Beidh Barra ar an bád leo amárach). The *Intonation* column provides the IViE labels at the phonological tier, followed by its respective pitch accent. The *Stress Pattern* column specifies the pitch accents of each recorded utterance (not necessarily reflecting the typical dialect intonation pattern).

SENTENCE 06 – Beidh Barra ar an bád leo amárach			
Affect	Repetition	Intonation	Stress Pattern
Neutral	Rep 1	H* [Ba], !H* [bád], !H*L [má]	%<Ba><bád><má>%
	Rep 2	H* [Ba], !H* [bád], !H*L [má]	%<Ba><bád><má>%
	Rep 3	H* [Ba], !H* [bád], !H*L [má]	%<Ba><bád><má>%
	Rep 4	H* [Ba], !H* [bád], !H*L [má]	%<Ba><bád><má>%
	Rep 5	H* [Ba], !H* [bád], !H*L [má]	%<Ba><bád><má>%
Angry	Rep 1	H*L [Ba], H*L [bád], H*L [má]	%<Ba><bád><má>%
	Rep 2	H*L [Ba], H*L [bád], H*L [má]	%<Ba><bád><má>%
	Rep 3	H*L [Ba], H*L [bád], H*L [má]	%<Ba><bád><má>%
	Rep 4	H*L [Ba], H*L [bád], H*L [má]	%<Ba><bád><má>%
	Rep 5	H*L [Ba], H*L [bád], H*L [má]	%<Ba><bád><má>%

Table A7 - Intonational Annotation of Sentence 7 (Beidh an aimsear go hálainn amárach).
The *Intonation* column provides the IViE labels at the phonological tier, followed by its respective pitch accent. The *Stress Pattern* column specifies the pitch accents of each recorded utterance (not necessarily reflecting the typical dialect intonation pattern).

SENTENCE 07 – Beidh an aimsear go hálainn amárach			
Affect	Repetition	Intonation	Stress Pattern
Neutral	Rep 1	H* [aim], !H*L [há], H*L [má]	%<aim><há><má>%
	Rep 2	H* [aim], !H*L [há], H*L [má]	%<aim><há><má>%
	Rep 3	H* [aim], !H*L [há], H*L [má]	%<aim><há><má>%
	Rep 4	H* [aim], !H*L [há], H*L [má]	%<aim><há><má>%
	Rep 5	H* [aim], !H*L [há], H*L [má]	%<aim><há><má>%
Angry	Rep 1	H*L [aim], H*L [há], H*L [má]	%<aim><há><má>%
	Rep 2	H*L [aim], H*L [há], H*L [má]	%<aim><há><má>%
	Rep 3	H*L [aim], H*L [há], H*L [má]	%<aim><há><má>%
	Rep 4	H*L [aim], H*L [há], H*L [má]	%<aim><há><má>%
	Rep 5	H*L [aim], H*L [há], H*L [má]	%<aim><há><má>%

Table A8 - Intonational Annotation of Sentence 8 (Beidh an aimsear ag báistí amárach).
The *Intonation* column provides the IViE labels at the phonological tier, followed by its respective pitch accent. The *Stress Pattern* column specifies the pitch accents of each recorded utterance (not necessarily reflecting the typical dialect intonation pattern).

SENTENCE 08 – Beidh an aimsear ag báistí amárach			
Affect	Repetition	Intonation	Stress Pattern
Neutral	Rep 1	H* [aim], !H* [báis], !H*L [má]	%<aim><báis><tí><má>%
	Rep 2	H* [aim], !H* [báis], !H*L [má]	%<aim><báis><tí><má>%
	Rep 3	H* [aim], !H* [báis], !H*L [má]	%<aim><báis><tí><má>%
	Rep 4	H* [aim], !H* [báis], !H*L [má]	%<aim><báis><tí><má>%
	Rep 5	H* [aim], !H* [báis], !H*L [má]	%<aim><báis><tí><má>%
Angry	Rep 1	H*L [aim], H*L [báis], H*% [má]	%<aim><báis><tí><má>%
	Rep 2	H*L [aim], H*L [báis], H*% [má]	%<aim><báis><tí><má>%
	Rep 3	H*L [aim], H*L [báis], H*% [má]	%<aim><báis><tí><má>%
	Rep 4	H*L [aim], H*L [báis], H*% [má]	%<aim><báis><tí><má>%

Appendix B

Stylization images for all affects and all manipulation types for f_0 , EE and RD, overlaid onto the contours of the recorded neutral baseline (manual analysis) and the VSG resynthesis output (from first set of manipulations based on production data).

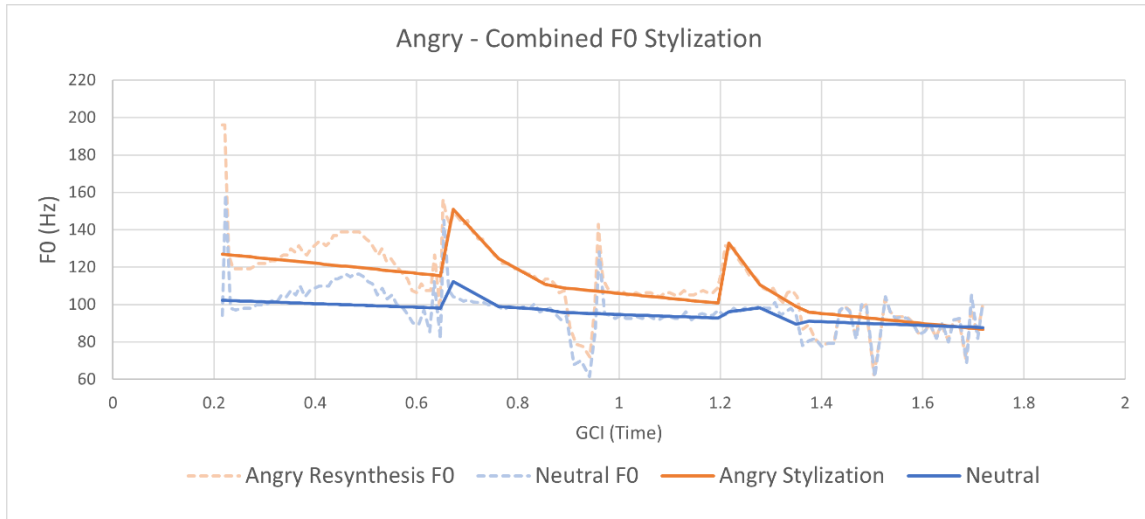


Figure B1 – F_0 stylization illustration for the *angry* combined resynthesized stimulus.

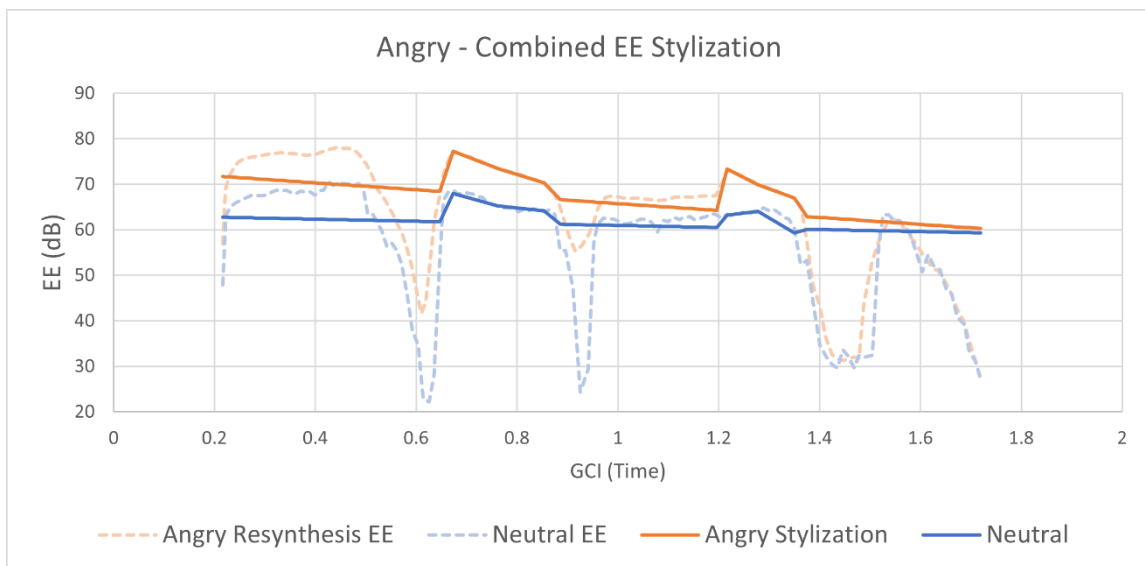


Figure B2 – EE stylization illustration for the *angry* combined resynthesized stimulus.

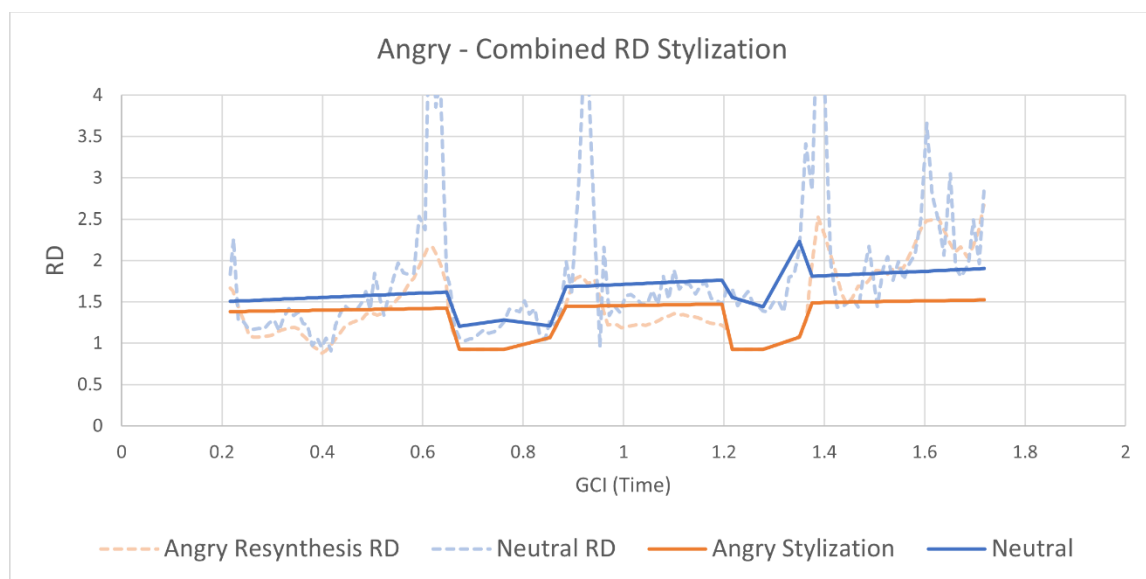


Figure B3 – RD stylization illustration for the *angry* combined resynthesized stimulus.

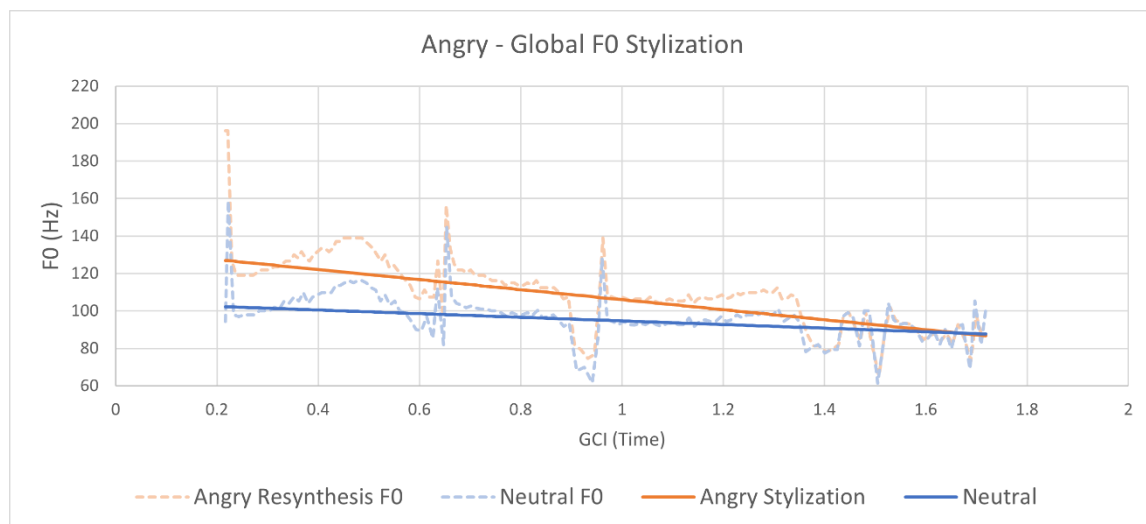


Figure B4 – F_0 stylization illustration for the *angry* global resynthesized stimulus.

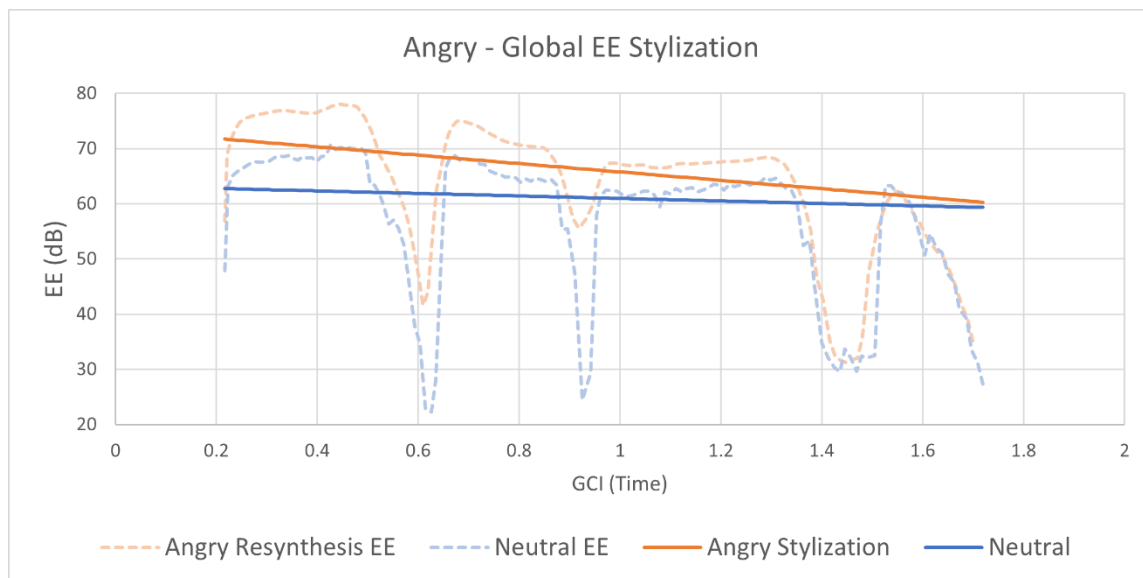


Figure B5 - EE stylization illustration for the *angry* global resynthesized stimulus.

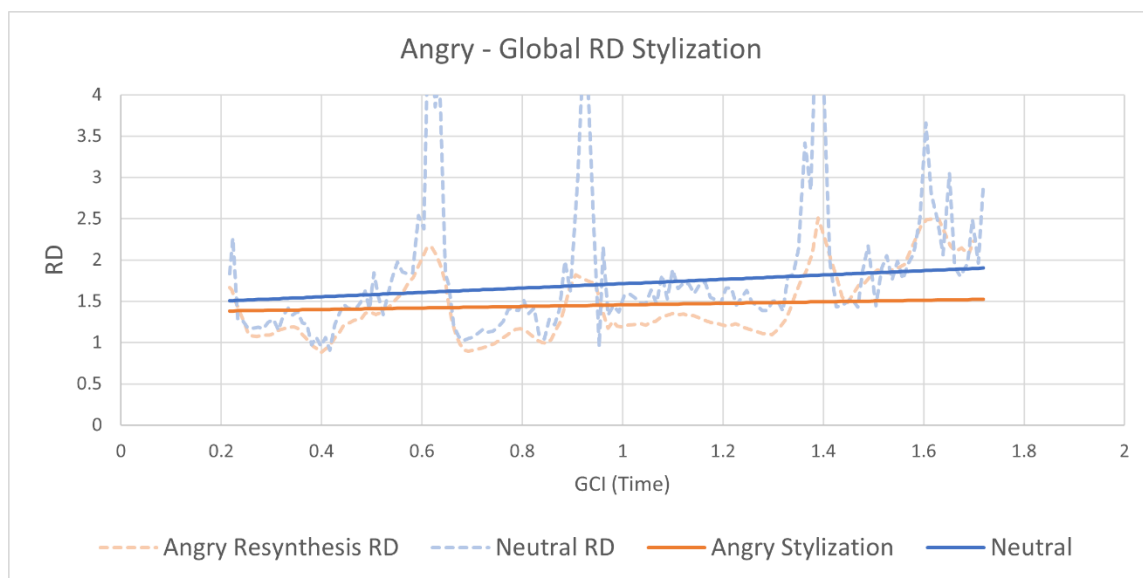


Figure B6 - RD stylization illustration for the *angry* global resynthesized stimulus.

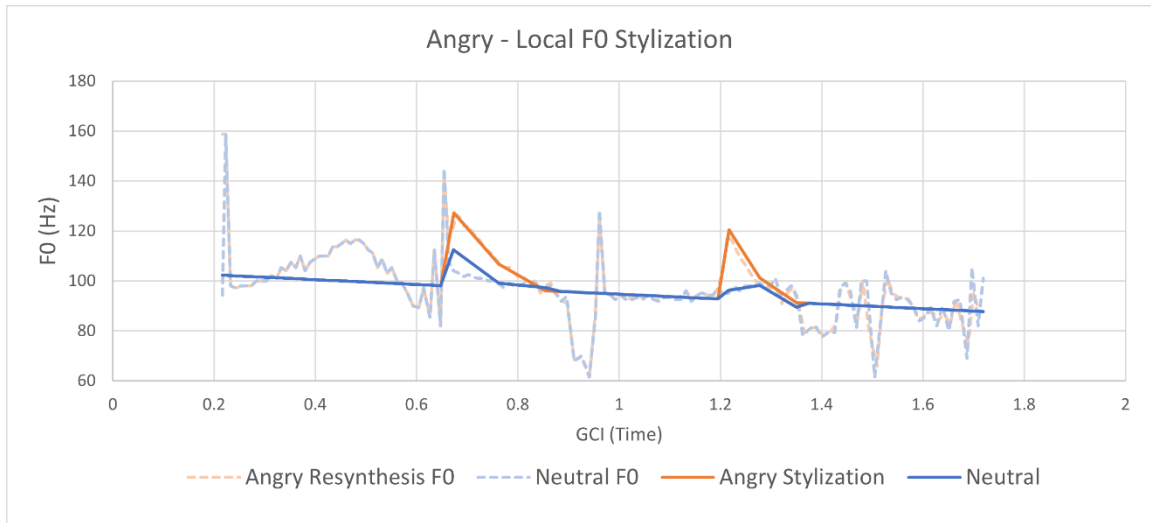


Figure B7 - F_0 stylization illustration for the *angry* local resynthesized stimulus.

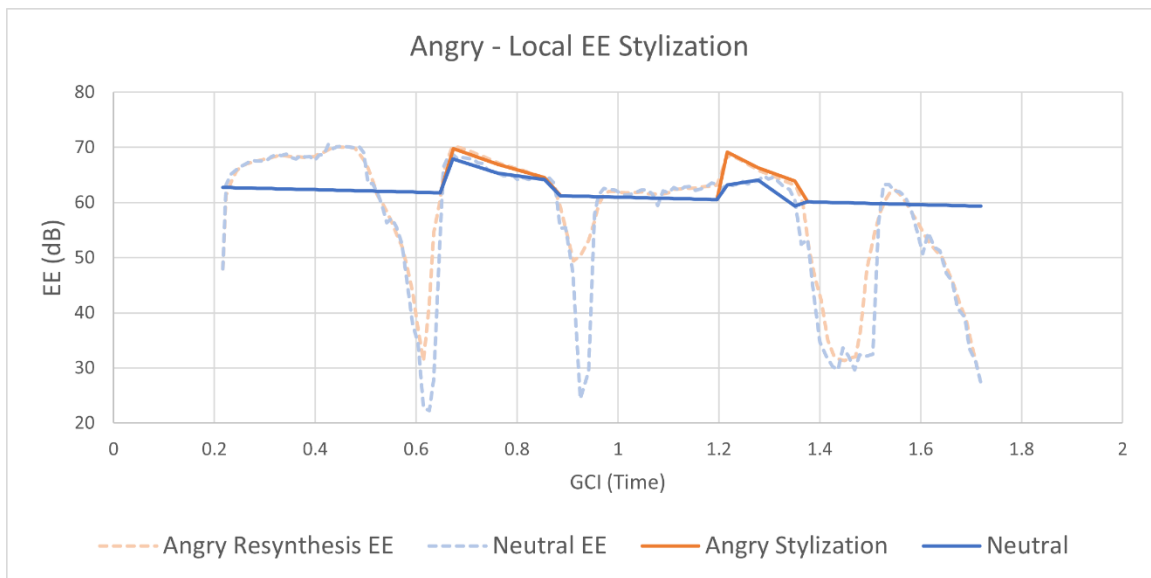


Figure B8 - EE stylization illustration for the *angry* local resynthesized stimulus.

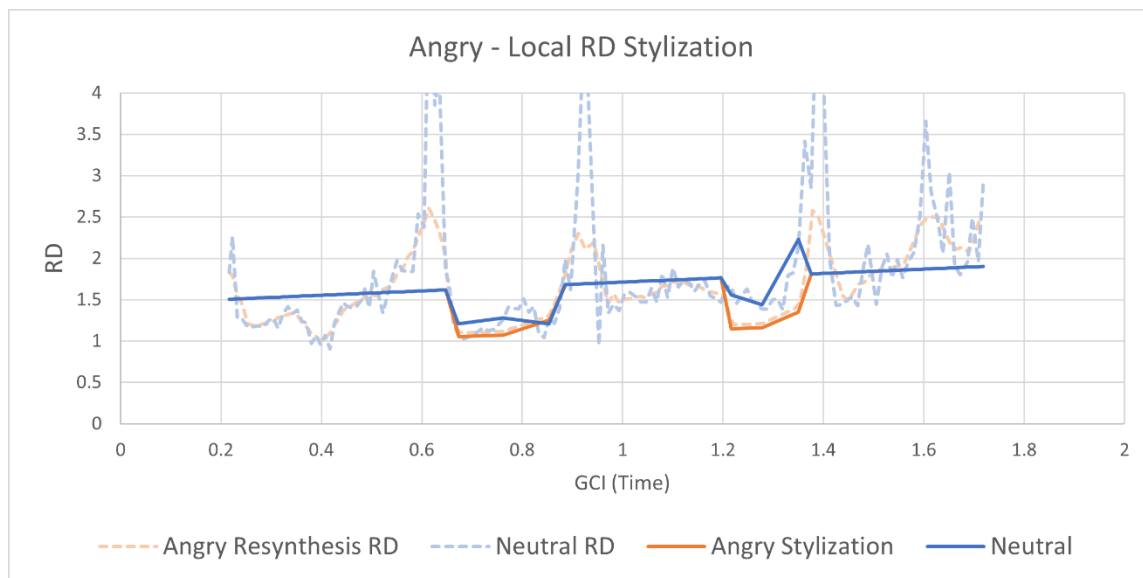


Figure B9 – RD stylization illustration for the *angry* local resynthesized stimulus.

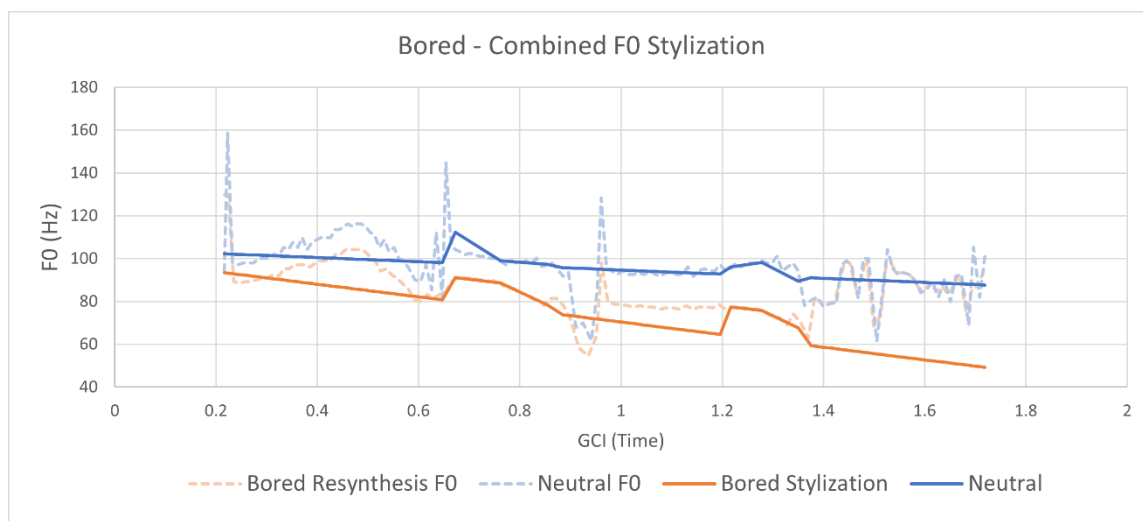


Figure B10 - F_0 stylization illustration for the *bored* combined resynthesized stimulus.

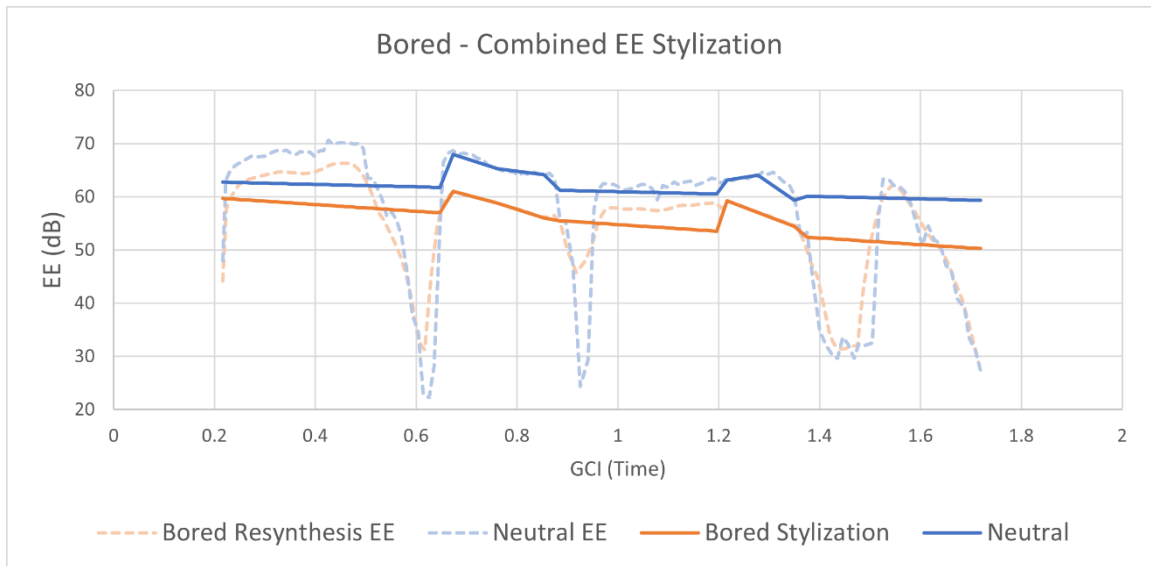


Figure B11 - EE stylization illustration for the *bored* combined resynthesized stimulus.

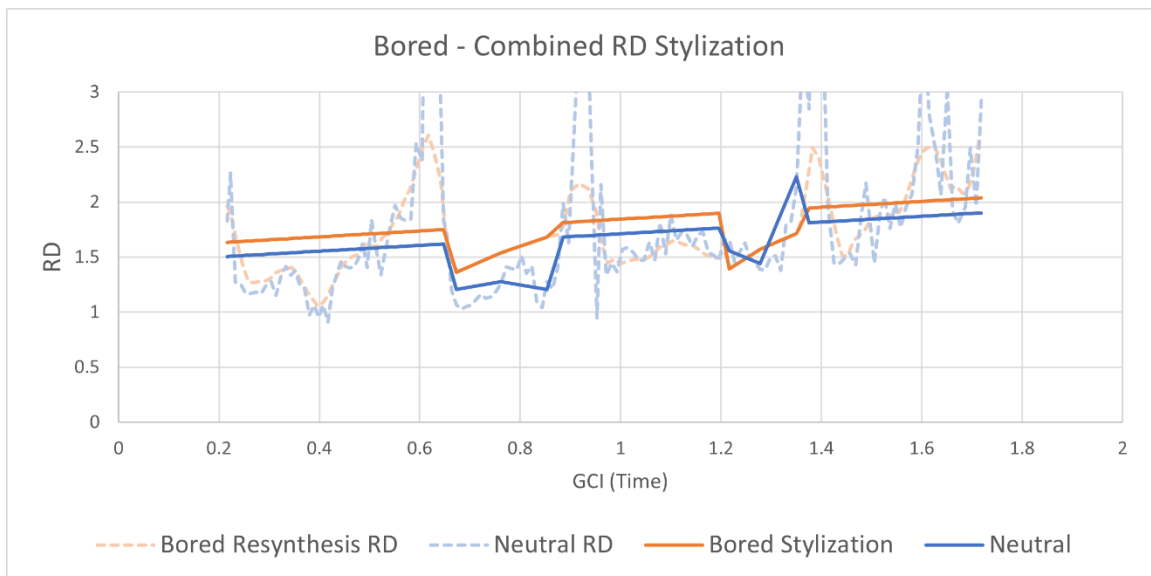


Figure B12 - RD stylization illustration for the *bored* combined resynthesized stimulus.

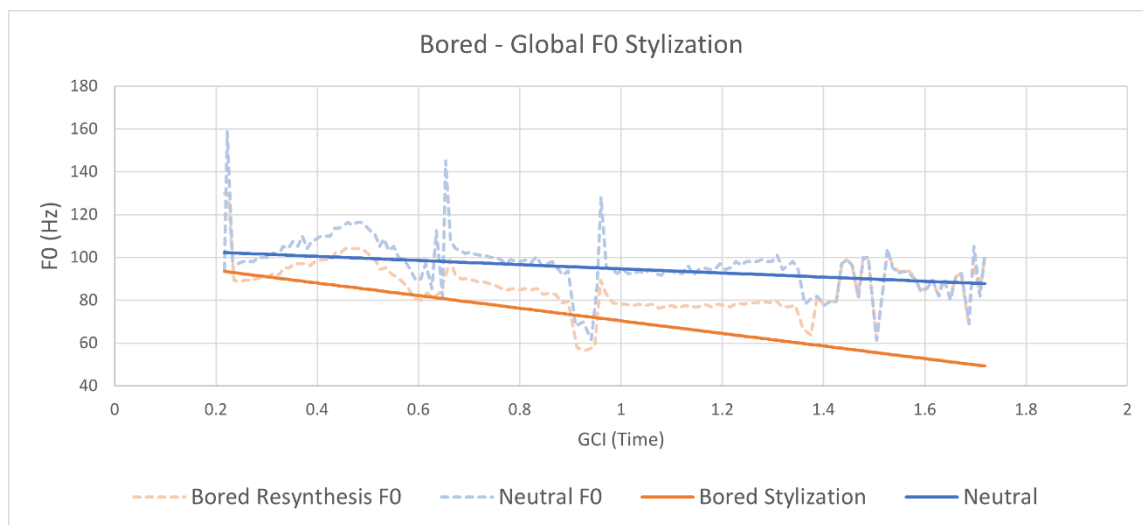


Figure B13 - F_0 stylization illustration for the *bored* global resynthesized stimulus.

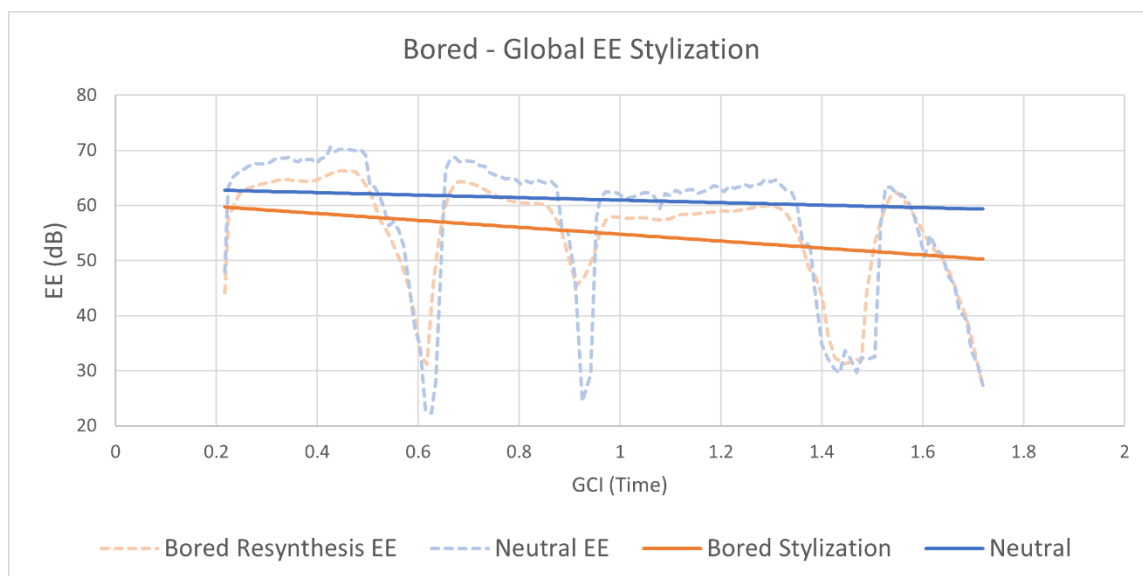


Figure B14 - EE stylization illustration for the *bored* global resynthesized stimulus.

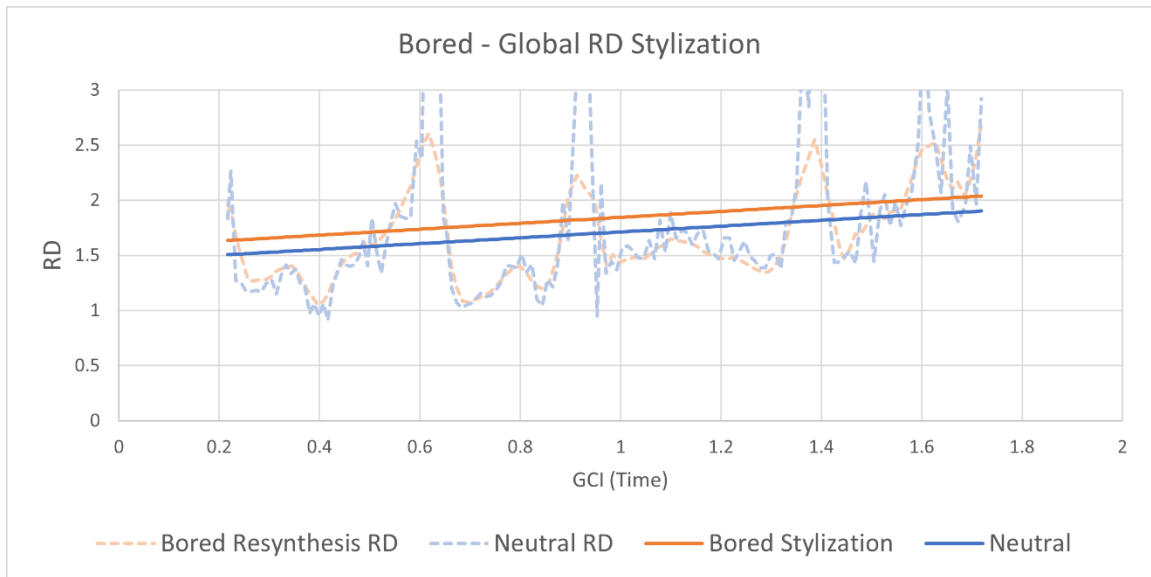


Figure B15 - RD stylization illustration for the *bored* global resynthesized stimulus.

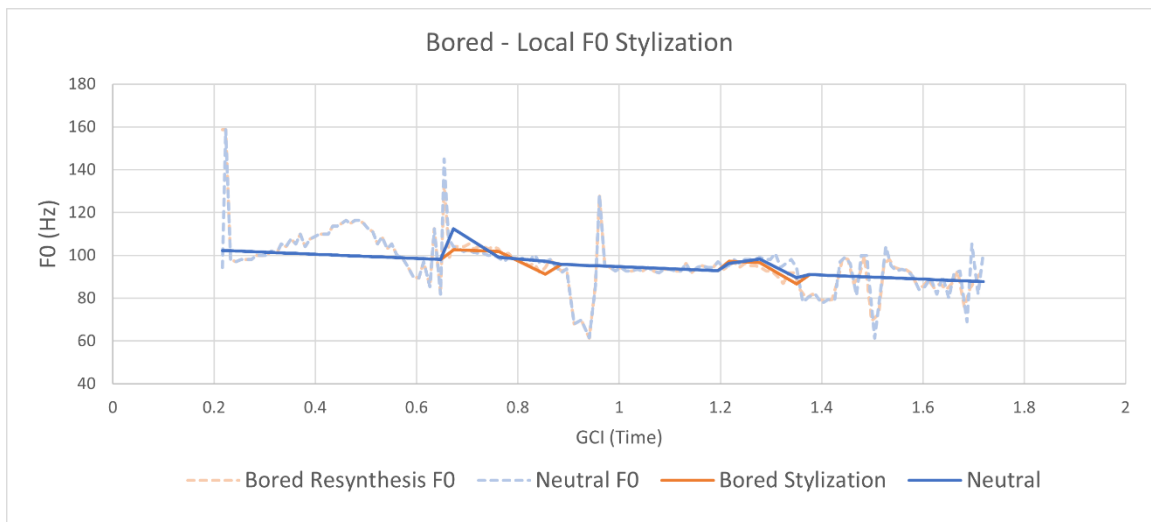


Figure B16 - F_0 stylization illustration for the *bored* local resynthesized stimulus.

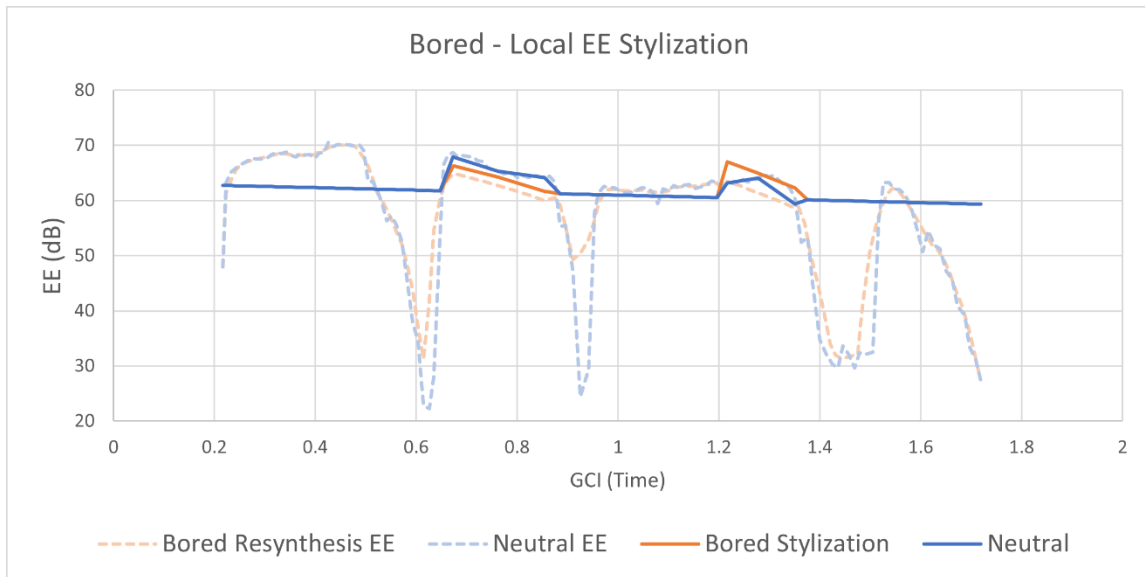


Figure B17 - EE stylization illustration for the *bored* local resynthesized stimulus.

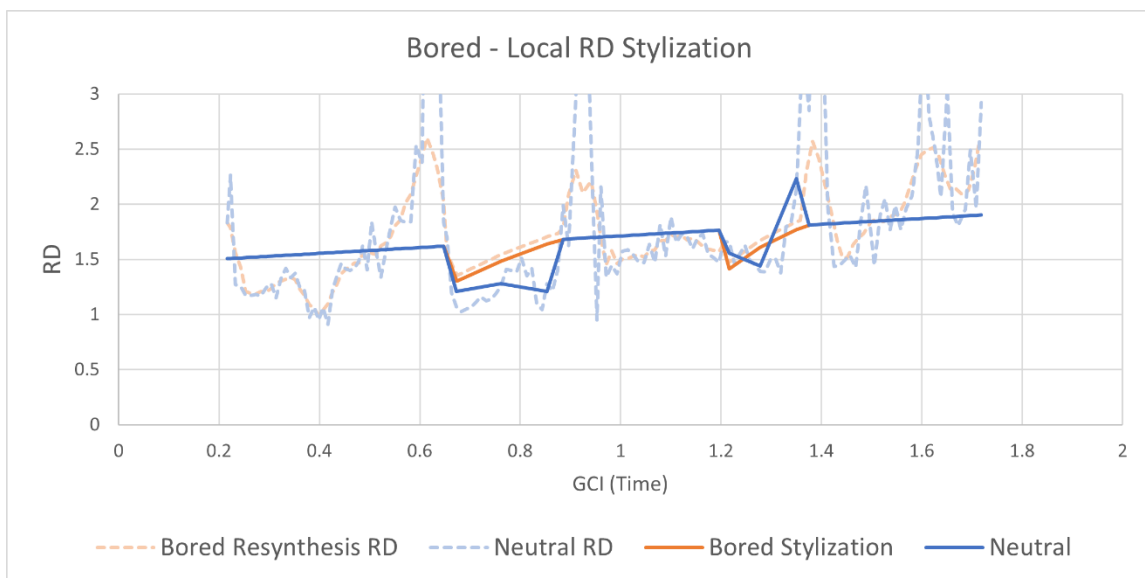


Figure B18 - F_0 RD stylization illustration for the *bored* local resynthesized stimulus.

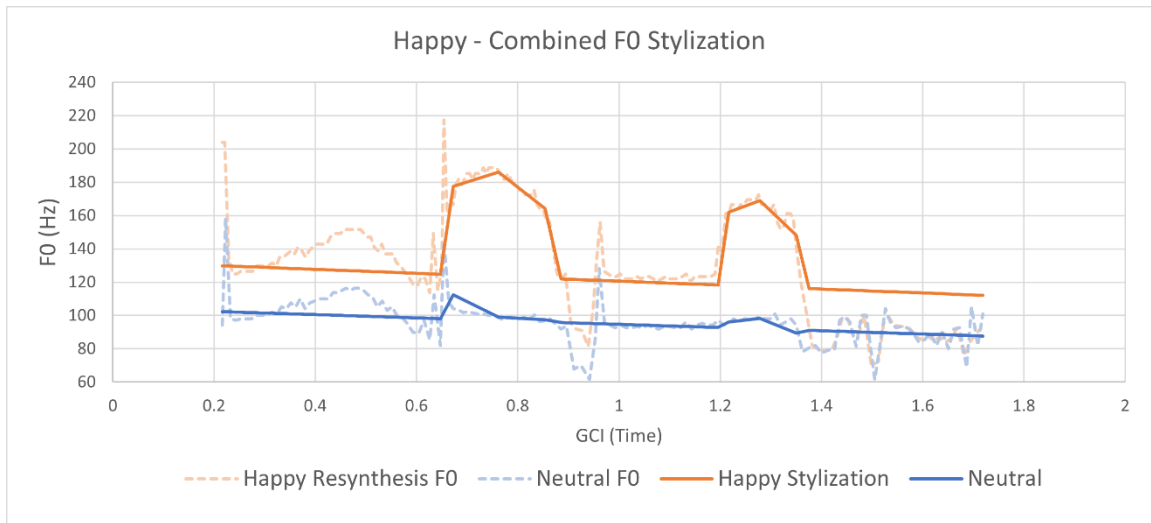


Figure B19 – F_0 stylization illustration for the *happy* combined resynthesized stimulus.

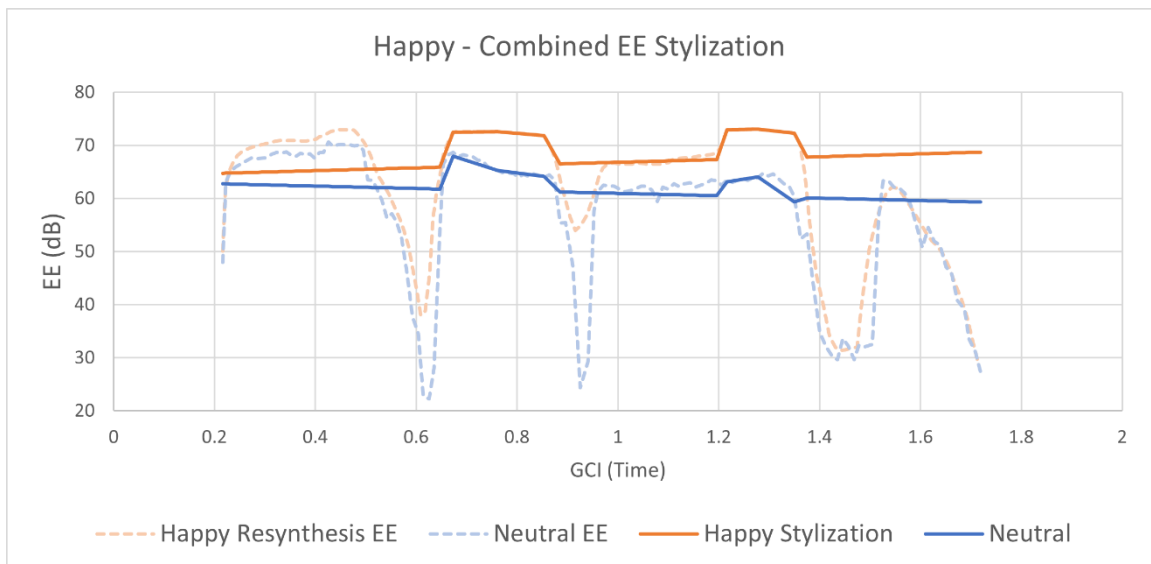


Figure B20 - EE stylization illustration for the *happy* combined resynthesized stimulus.

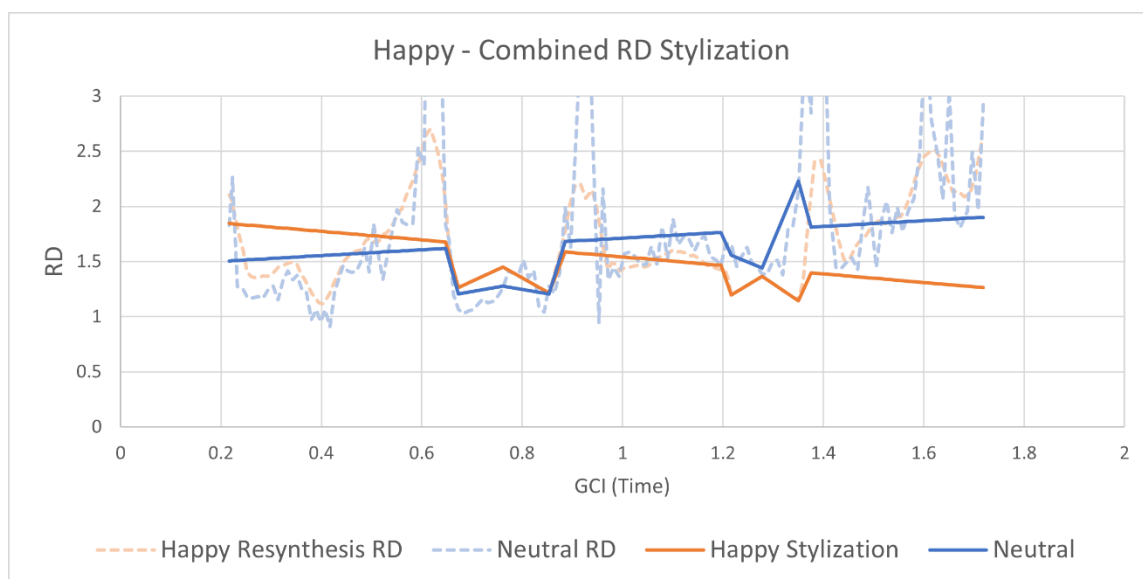


Figure B21 - RD stylization illustration for the *happy* combined resynthesized stimulus.

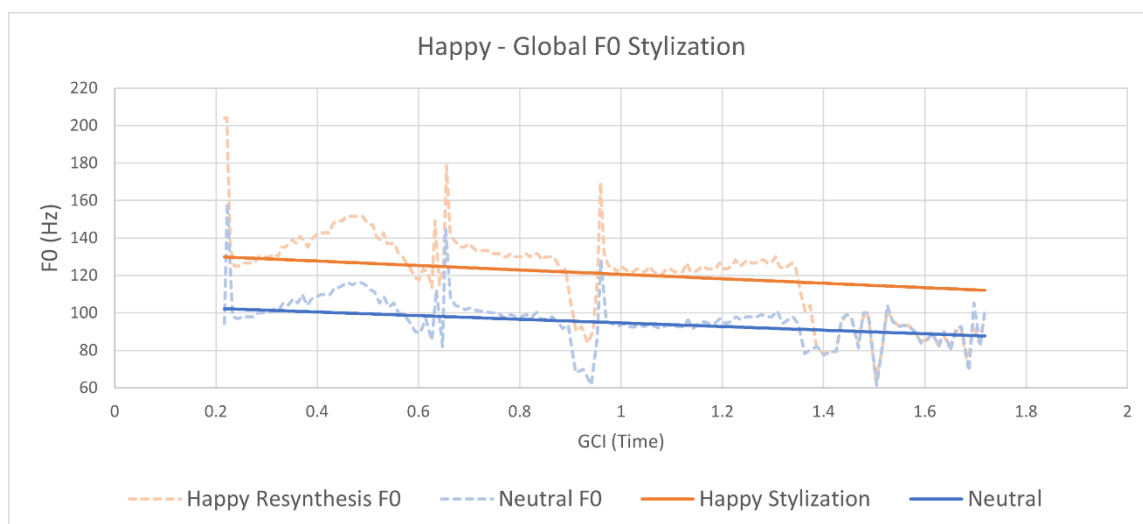


Figure B22 - F_0 stylization illustration for the *happy* global resynthesized stimulus.

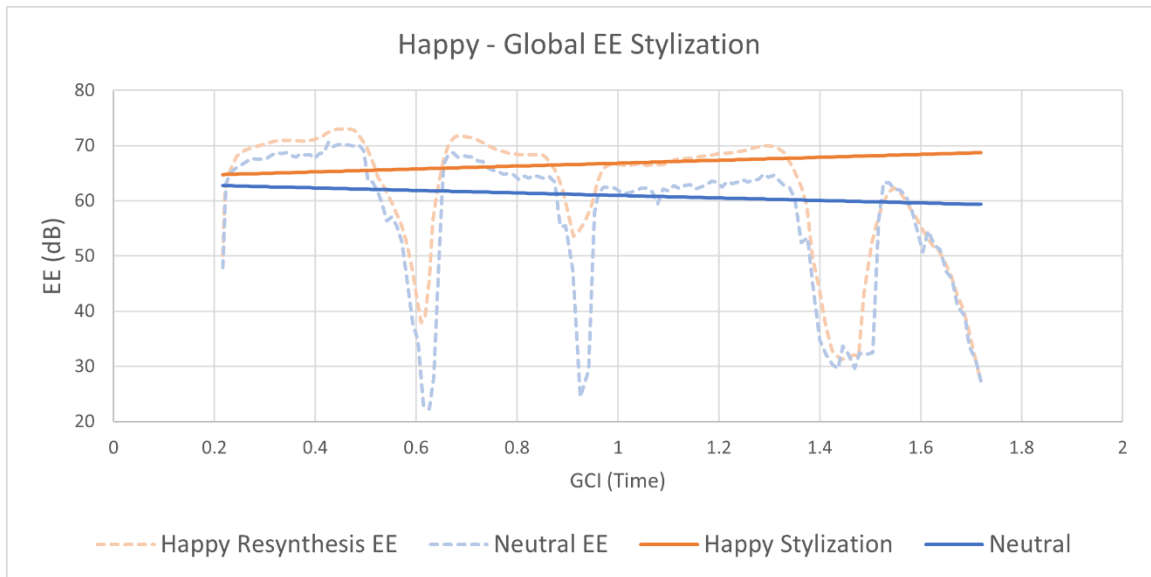


Figure B23 - EE stylization illustration for the *happy* global resynthesized stimulus.

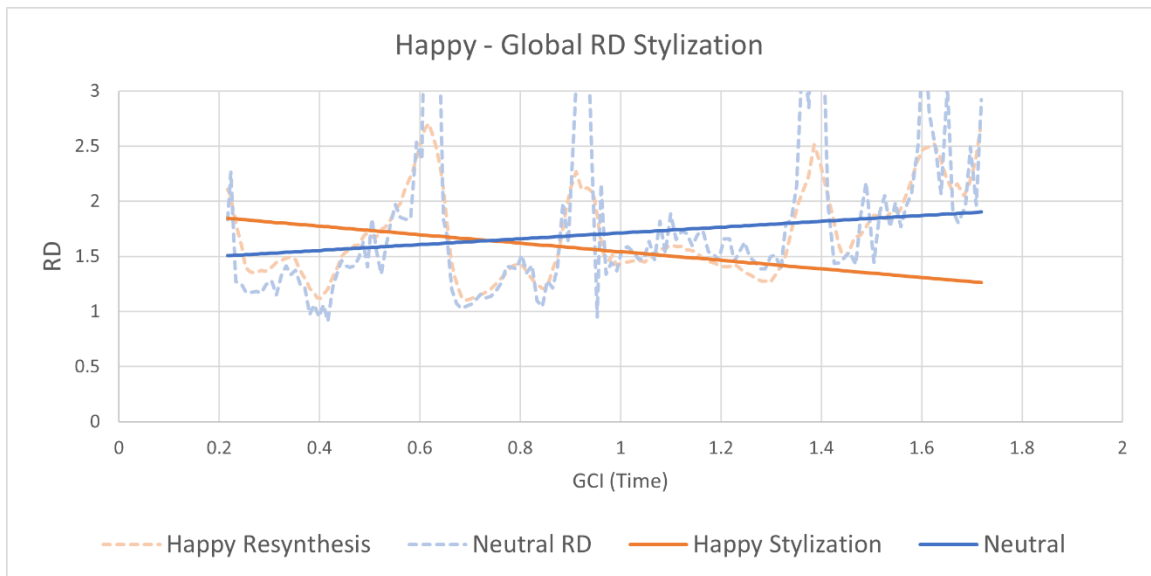


Figure B24 - RD stylization illustration for the *happy* global resynthesized stimulus.

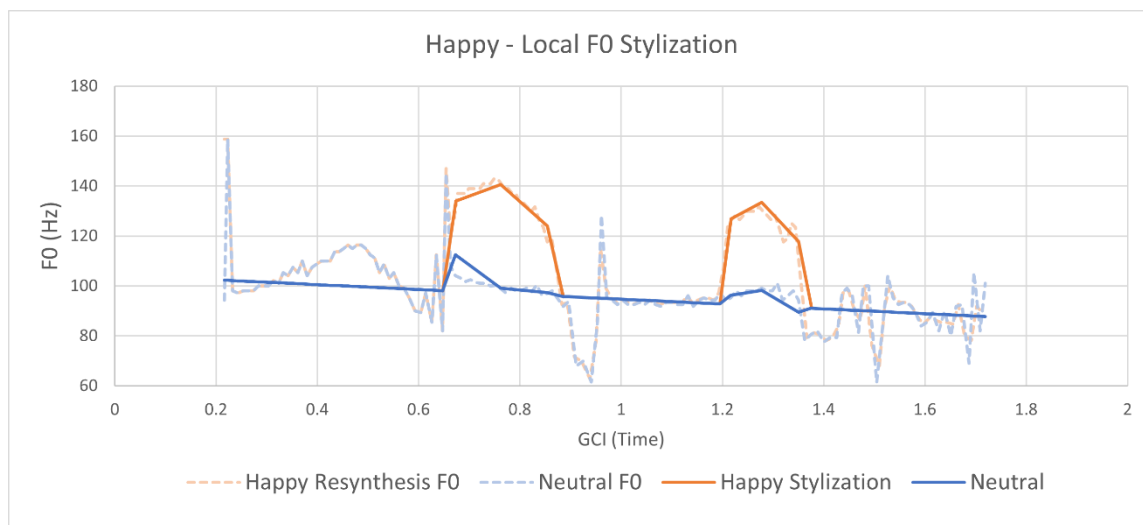


Figure B25 - F_0 stylization illustration for the *happy* local resynthesized stimulus.

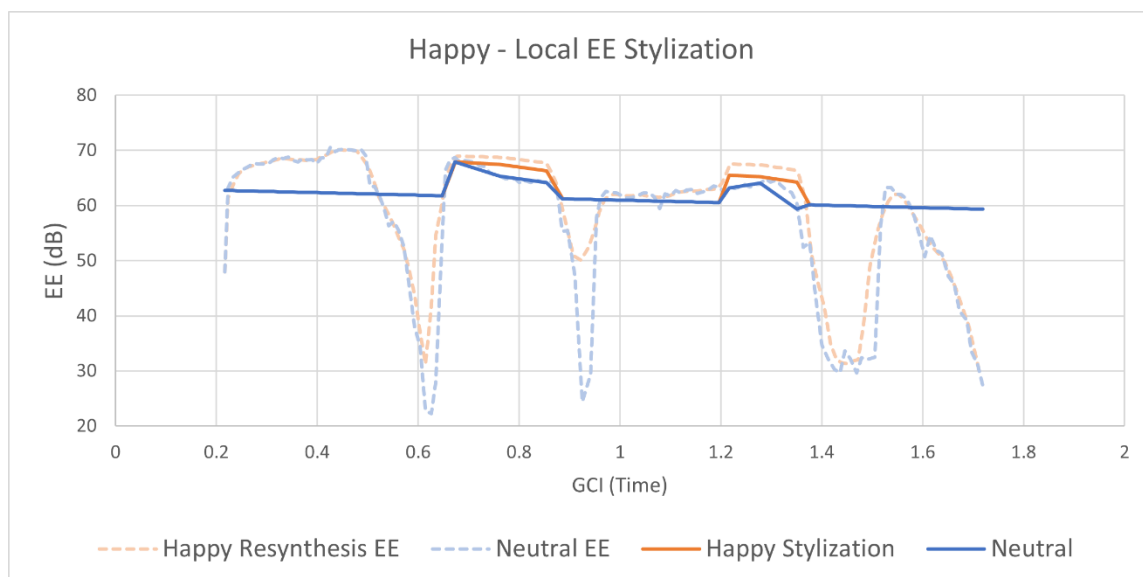


Figure B26 - EE stylization illustration for the *happy* local resynthesized stimulus.

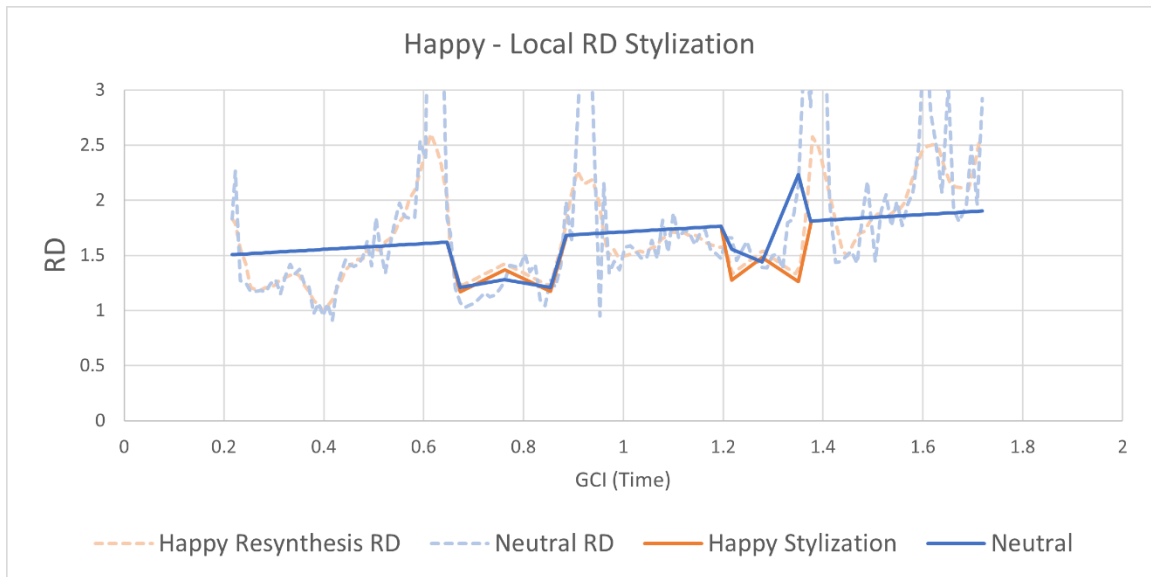


Figure B27 - RD stylization illustration for the *happy* local resynthesized stimulus.

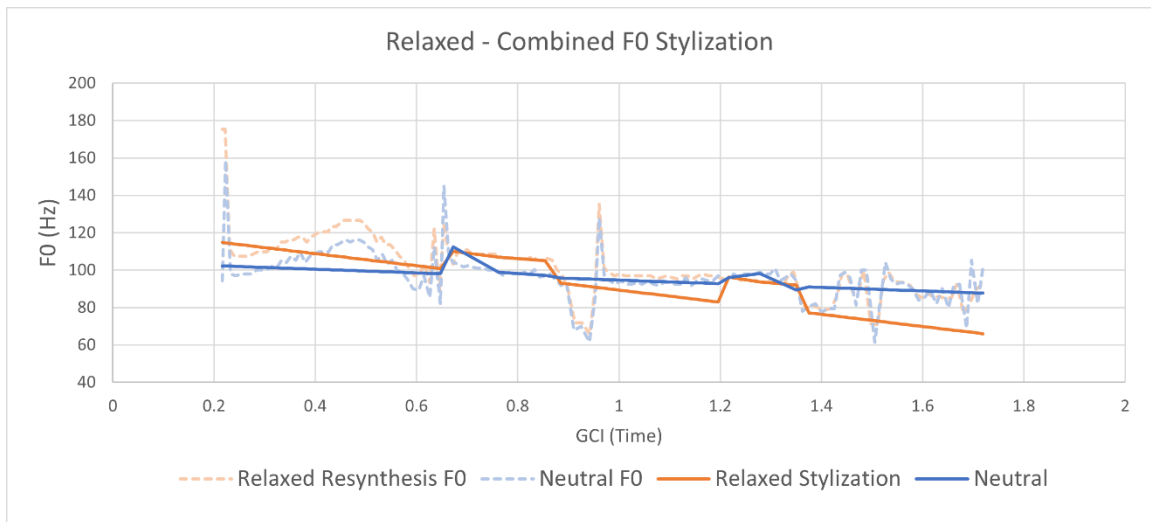


Figure B28 - F_0 stylization illustration for the *relaxed* combined resynthesized stimulus.

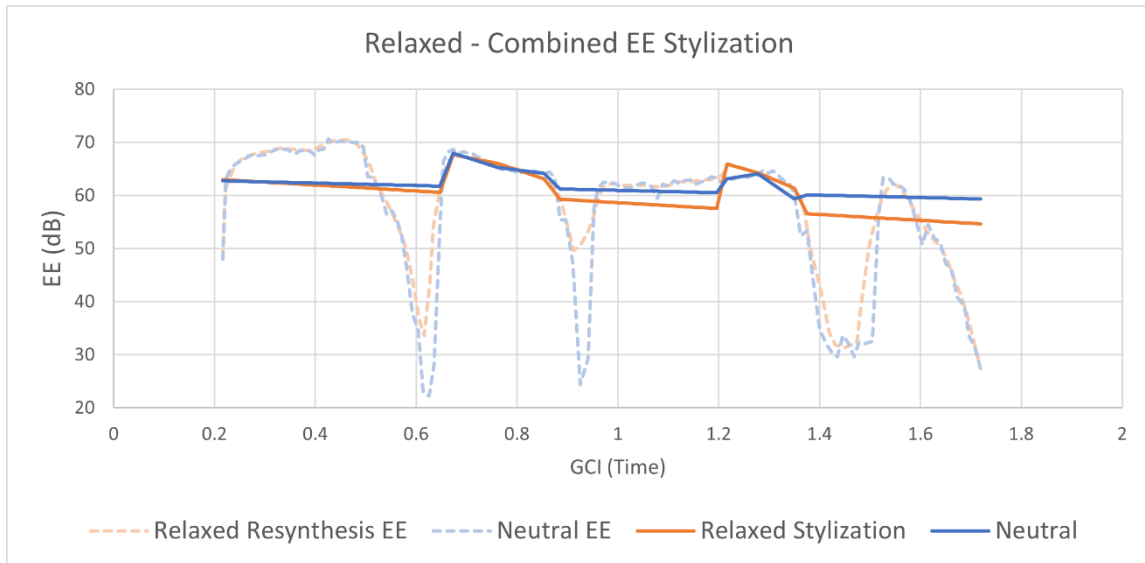


Figure B29 - EE stylization illustration for the *relaxed* combined resynthesized stimulus.

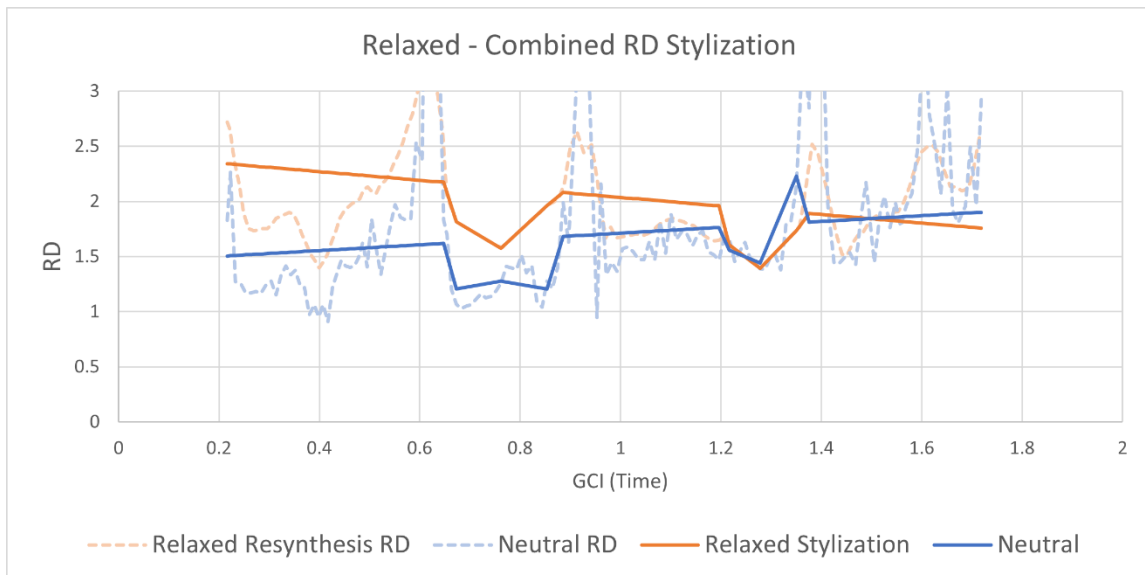


Figure B30 - RD stylization illustration for the *relaxed* combined resynthesized stimulus.

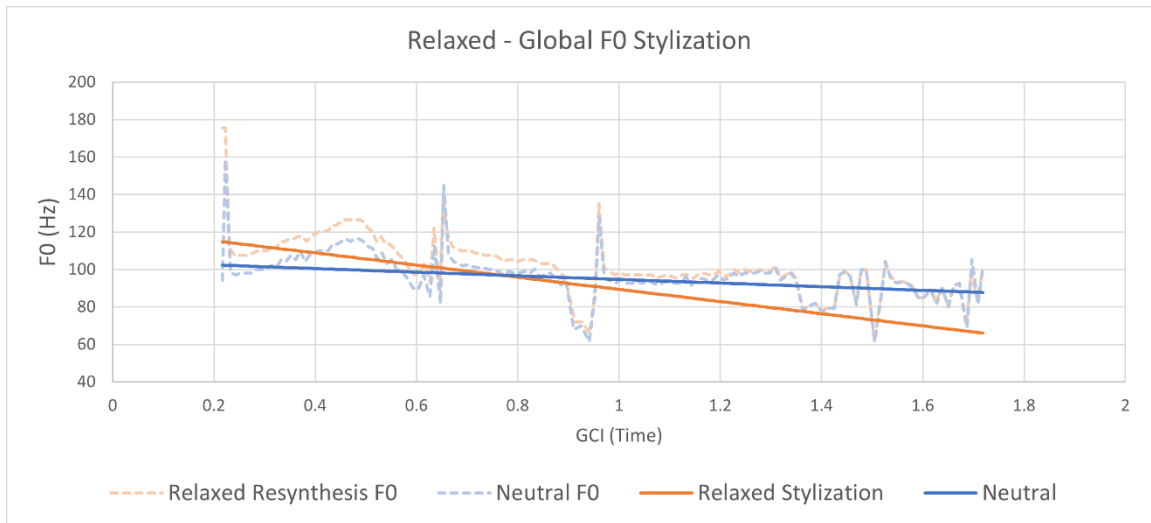


Figure B31 - F_0 stylization illustration for the *relaxed* global resynthesized stimulus.

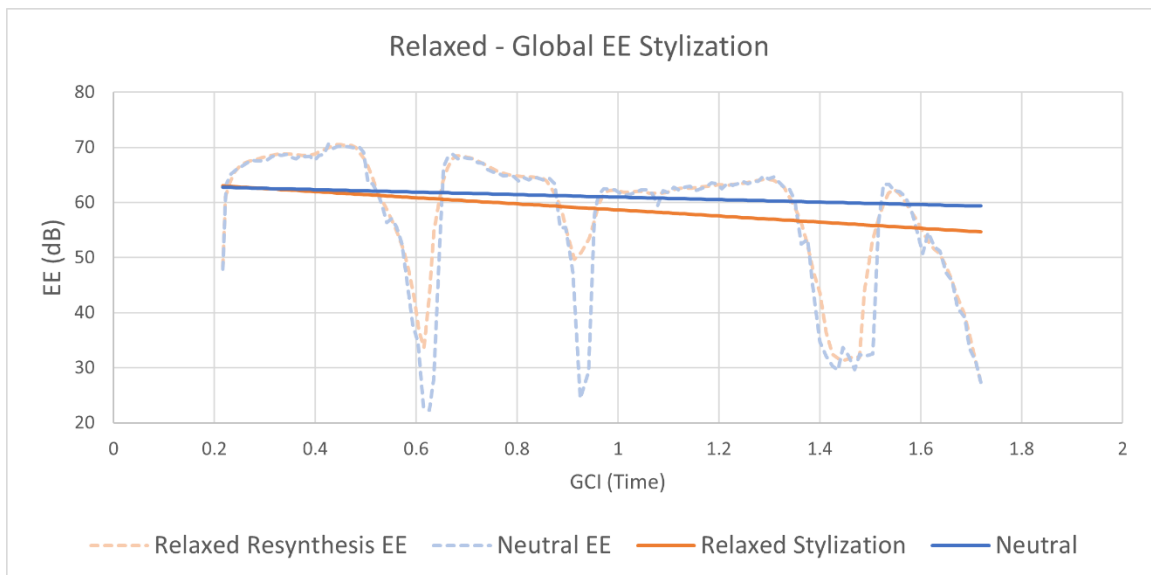


Figure B32 - EE stylization illustration for the *relaxed* global resynthesized stimulus.

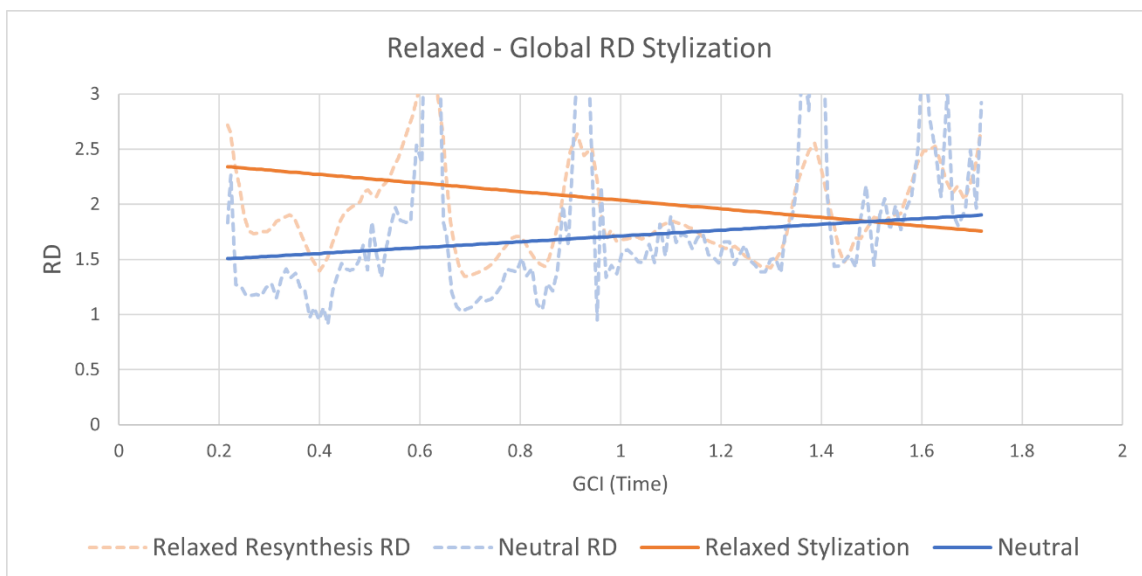


Figure B33 - RD stylization illustration for the *relaxed* global resynthesized stimulus.

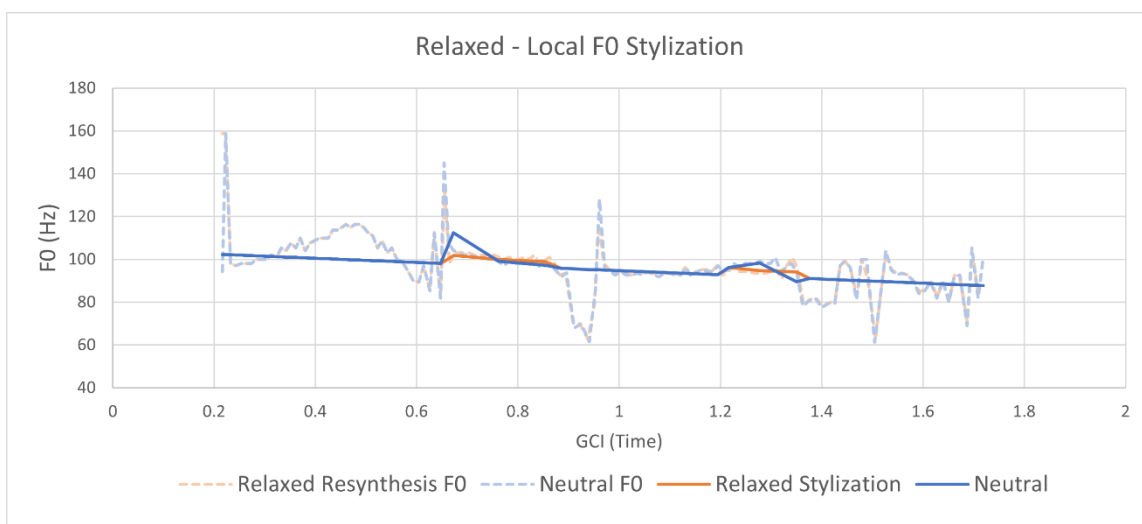


Figure B34 - F_0 stylization illustration for the *relaxed* local resynthesized stimulus.

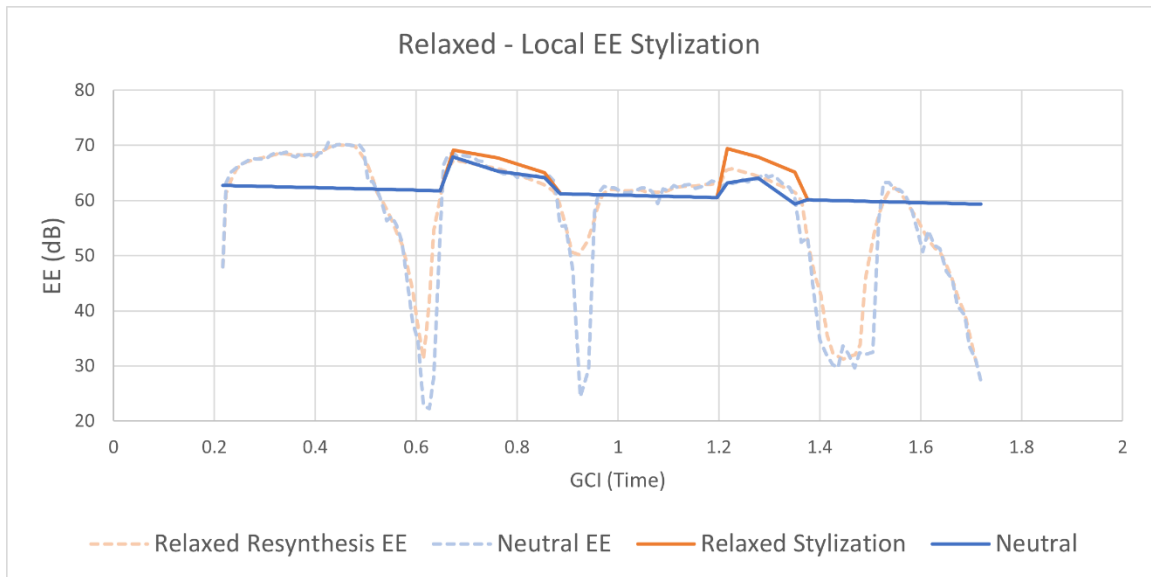


Figure B35 - EE stylization illustration for the *relaxed* combined resynthesized stimulus.

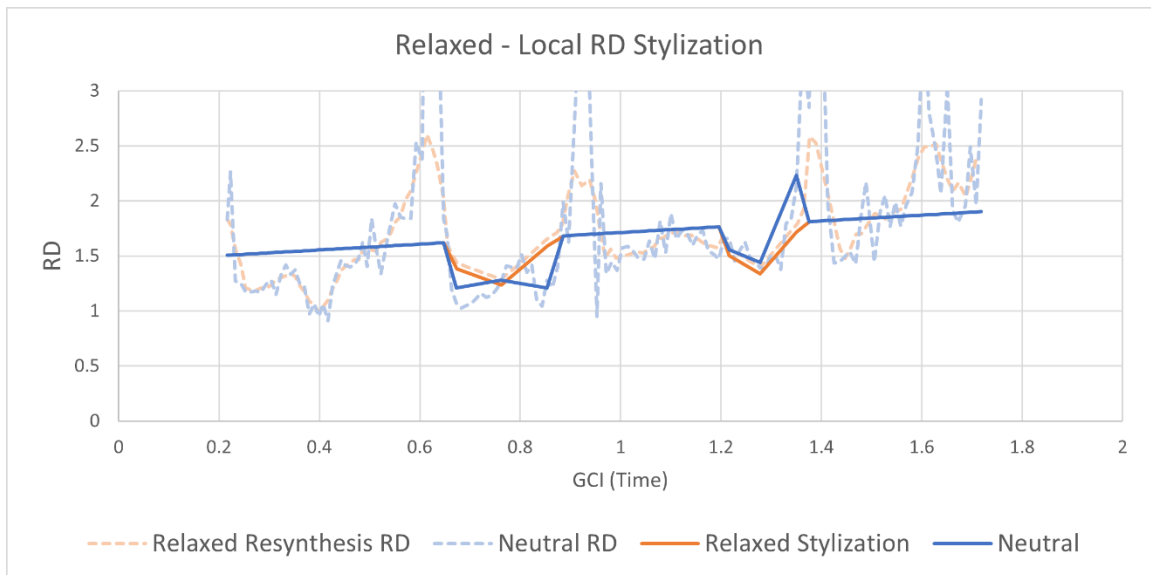


Figure B36 - RD stylization illustration for the *relaxed* local resynthesized stimulus.

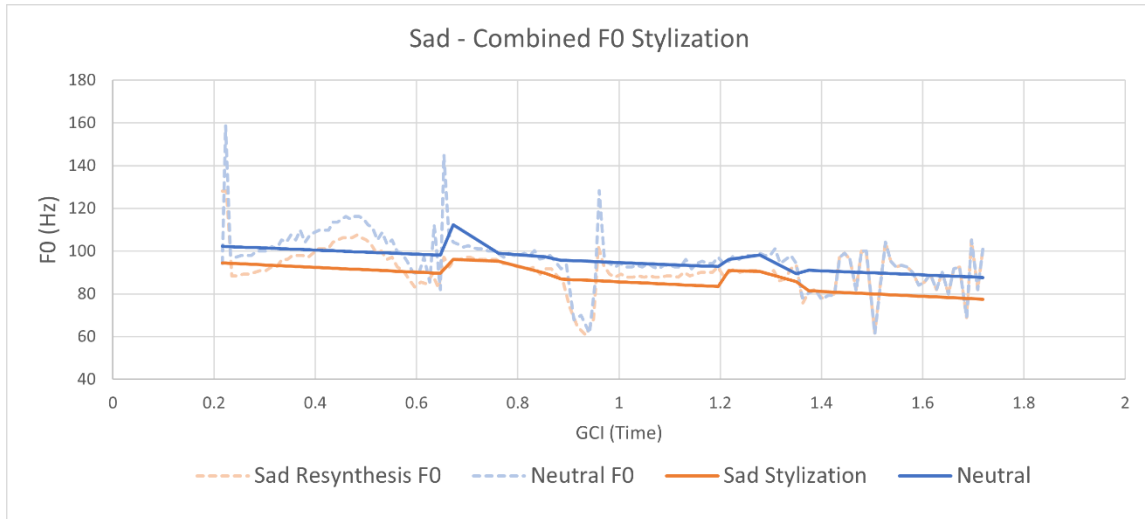


Figure B37 - F_0 stylization illustration for the *sad* combined resynthesized stimulus.

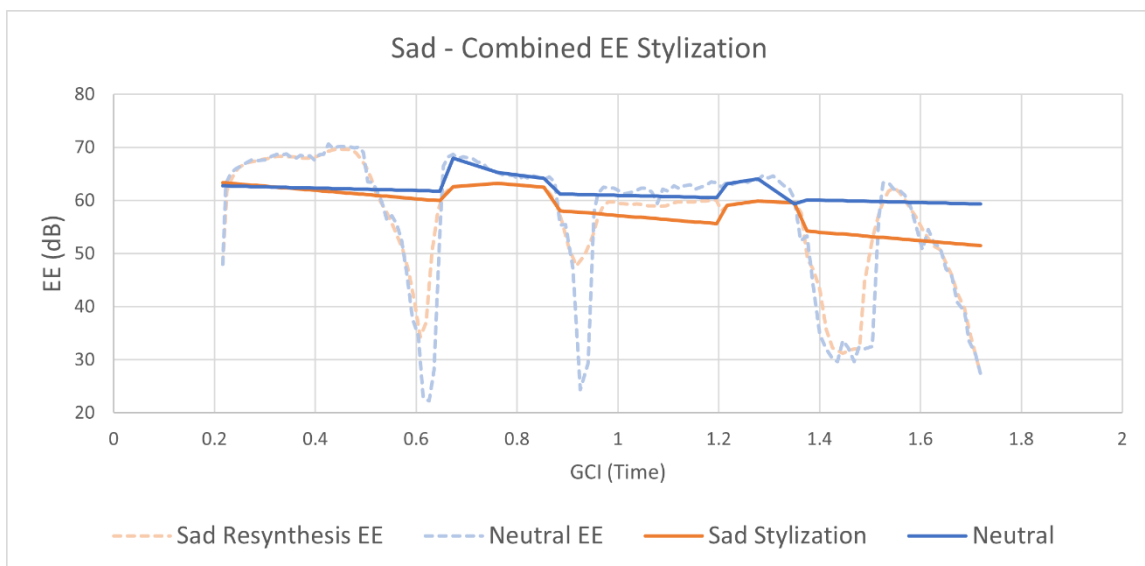


Figure B38 - EE stylization illustration for the *sad* combined resynthesized stimulus.

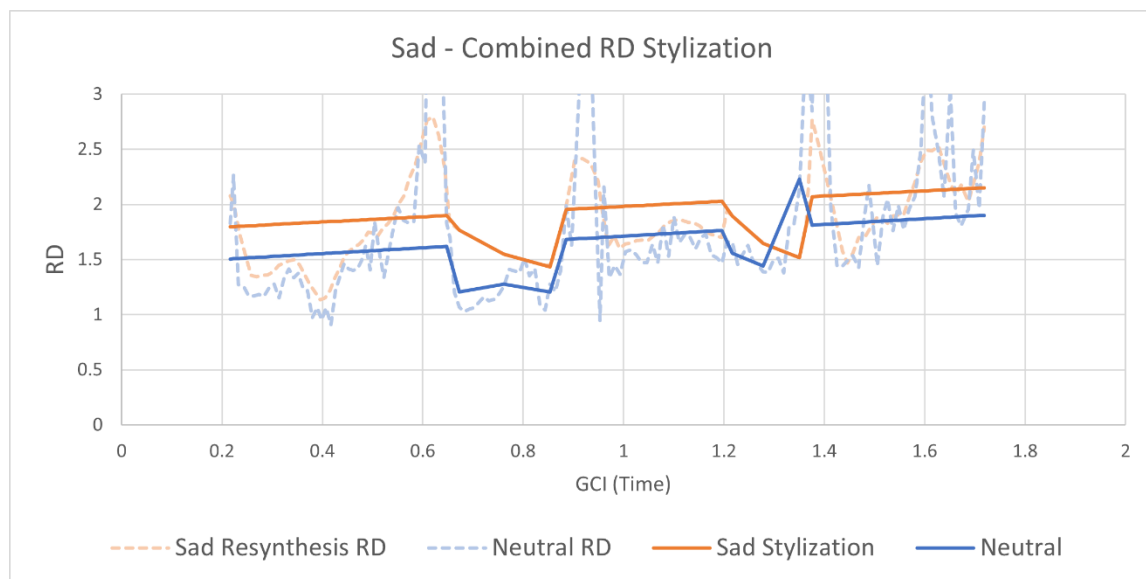


Figure B39 - RD stylization illustration for the *sad* combined resynthesized stimulus.

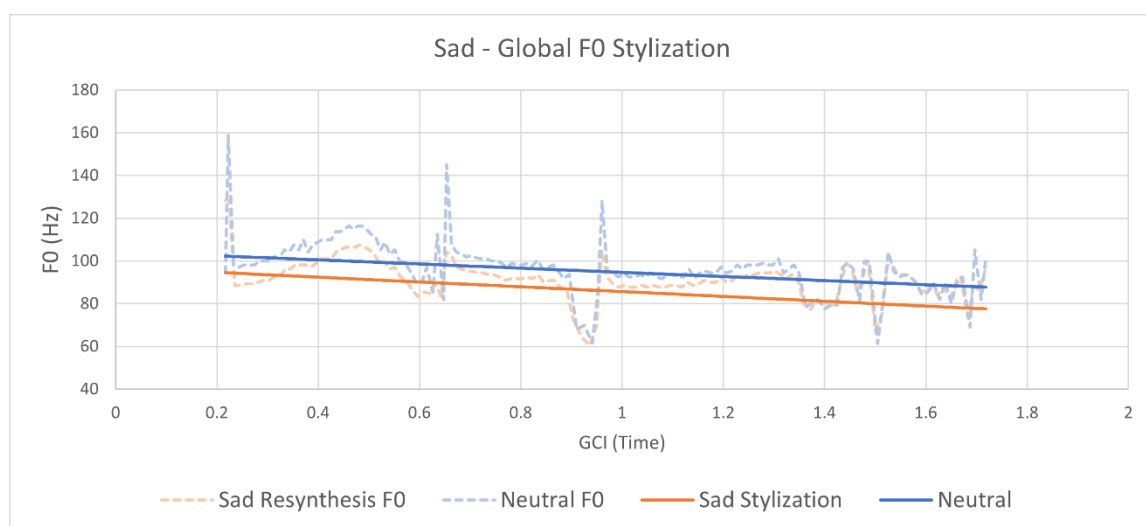


Figure B40 - F_0 stylization illustration for the *sad* global resynthesized stimulus.

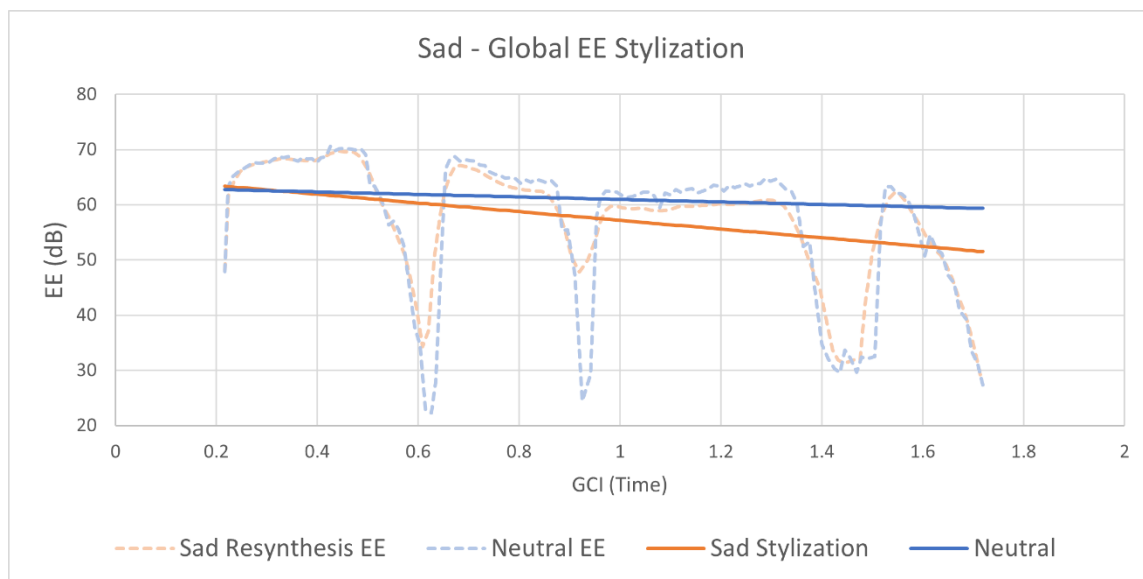


Figure B41 - EE stylization illustration for the *sad* global resynthesized stimulus.

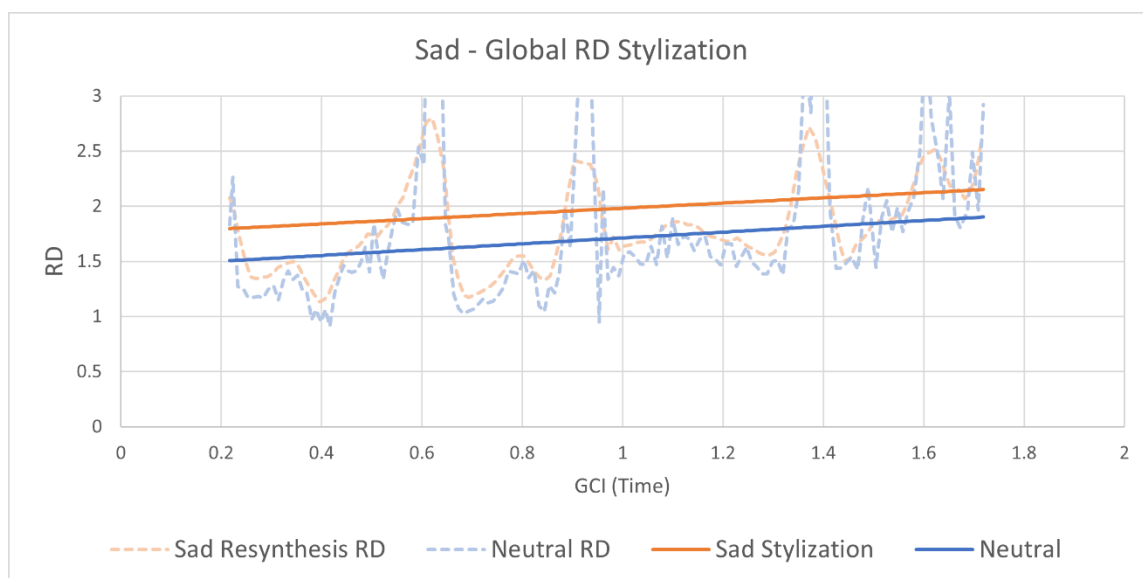


Figure B42 - RD stylization illustration for the *sad* global resynthesized stimulus.

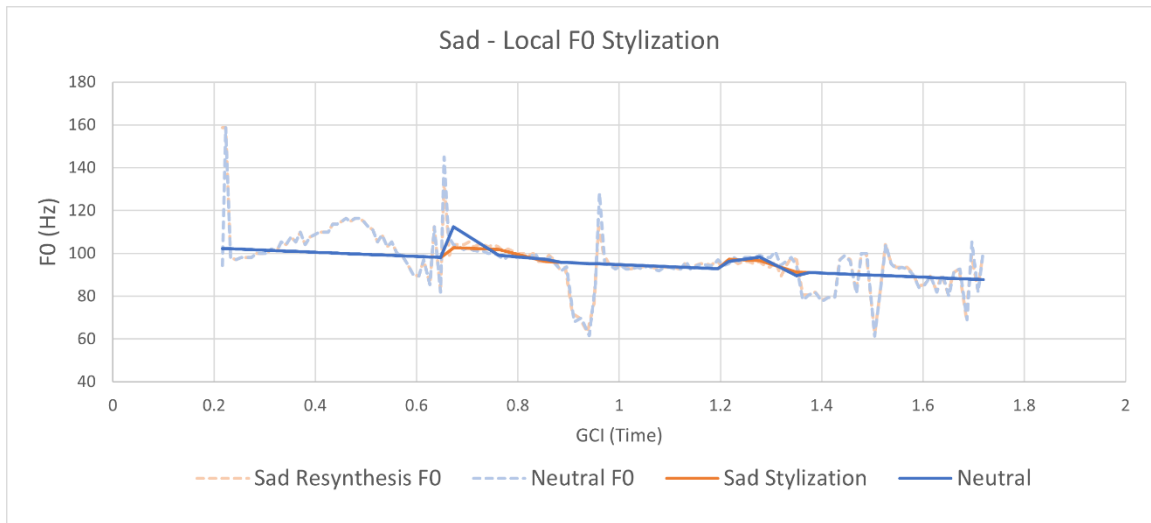


Figure B43 - F_0 stylization illustration for the *sad* local resynthesized stimulus.

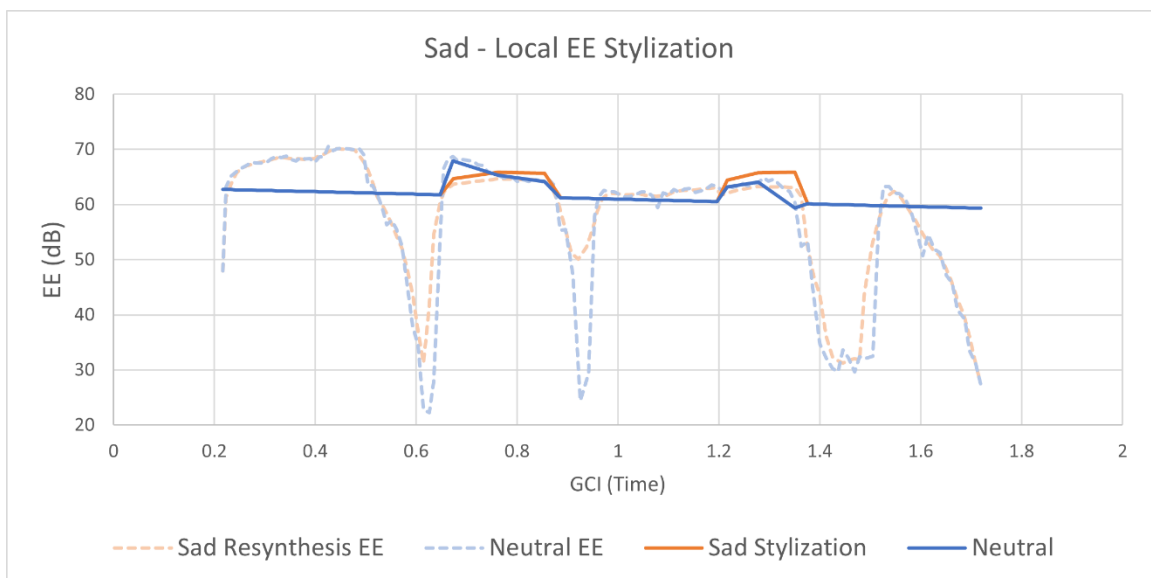


Figure B44 - EE stylization illustration for the *sad* local resynthesized stimulus.

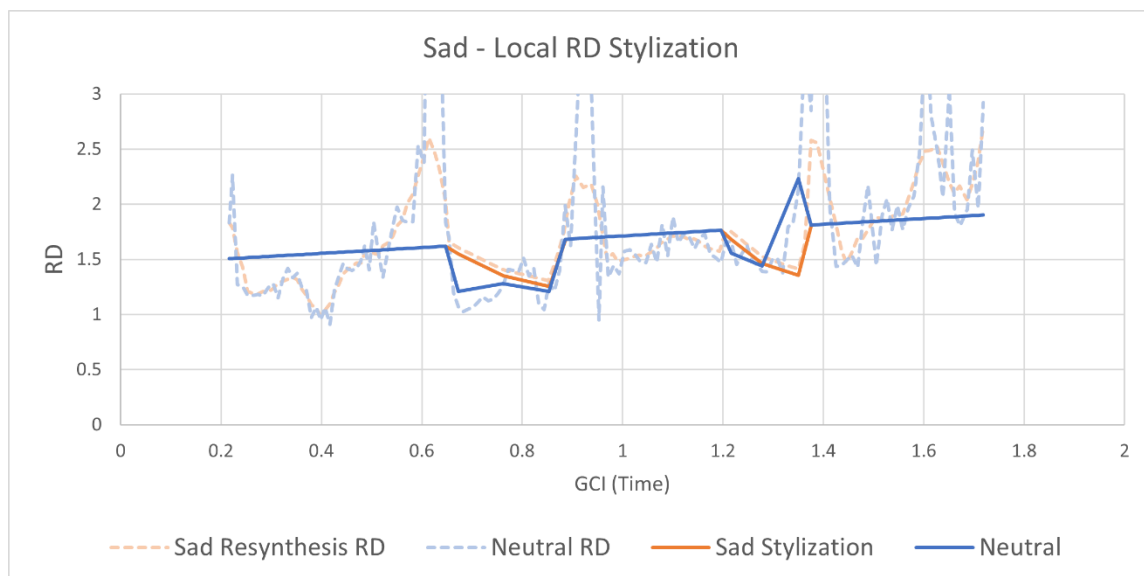


Figure B45 - RD stylization illustration for the *sad* local resynthesized stimulus.

Appendix C

All voice source parameter shifts for all affects for the recorded neutral baseline experiment.

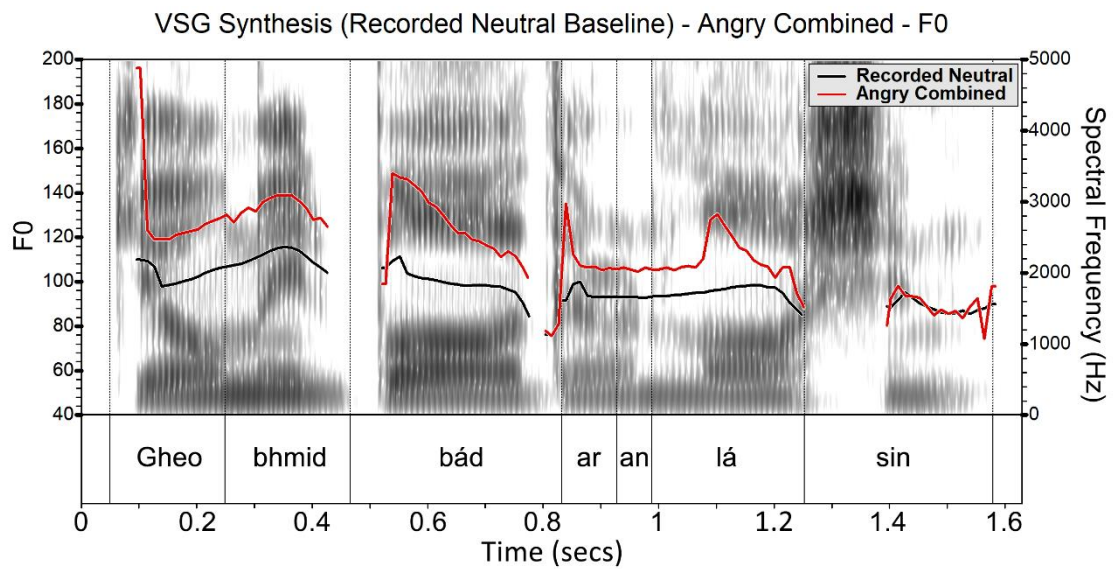


Figure C1 - F_0 contour of the combined modification of the recorded neutral stimulus to the *angry* target affect.

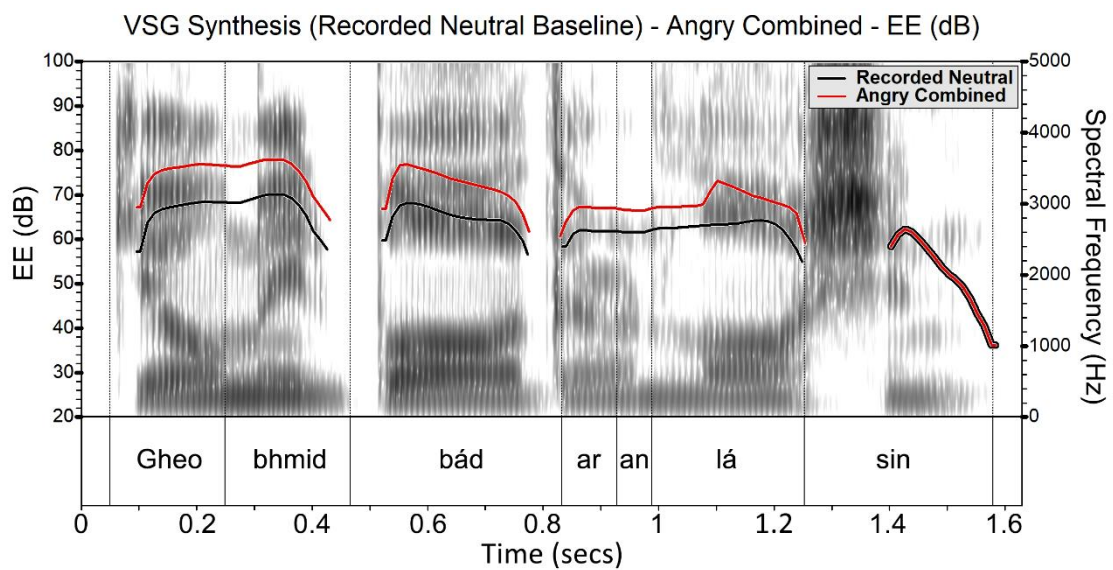


Figure C2 - EE contour of the combined modification of the recorded neutral stimulus to the *angry* target affect.

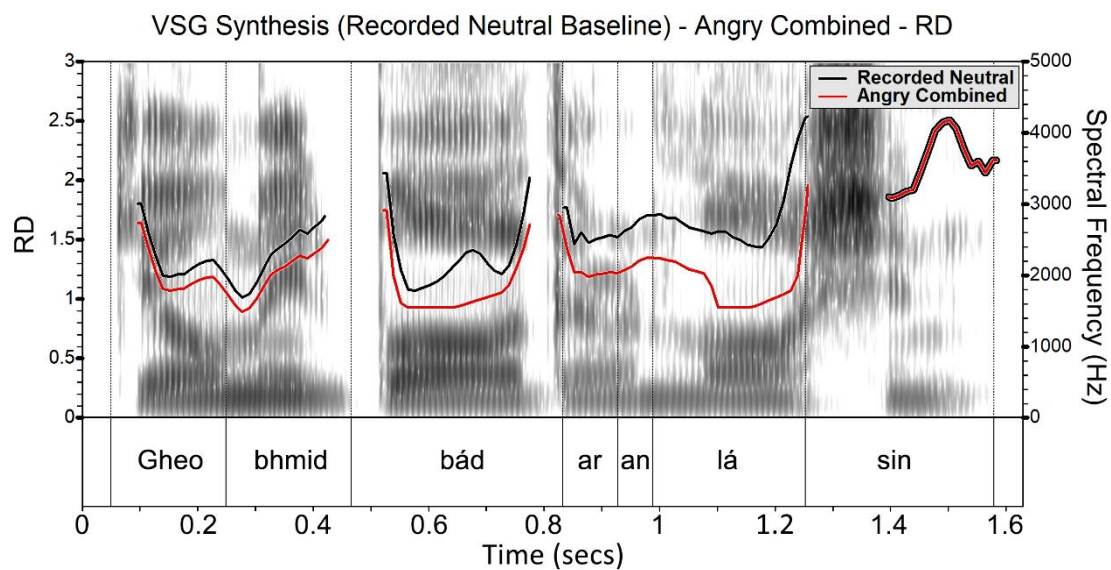


Figure C3 - RD contour of the combined modification of the recorded neutral stimulus to the *angry* target affect.

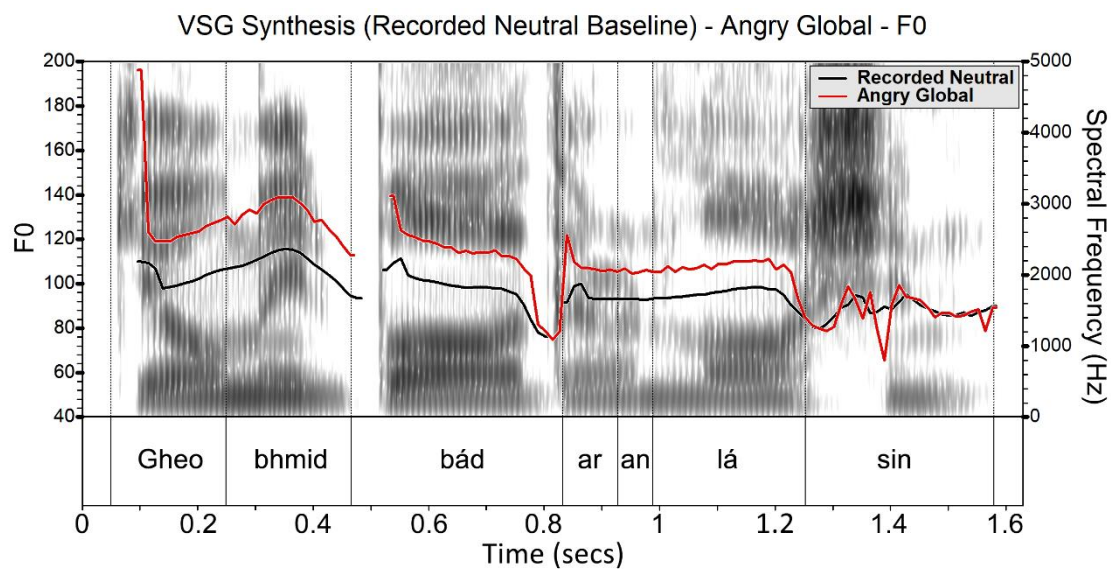


Figure C4 - F₀ contour of the global modification of the recorded neutral stimulus to the *angry* target affect.

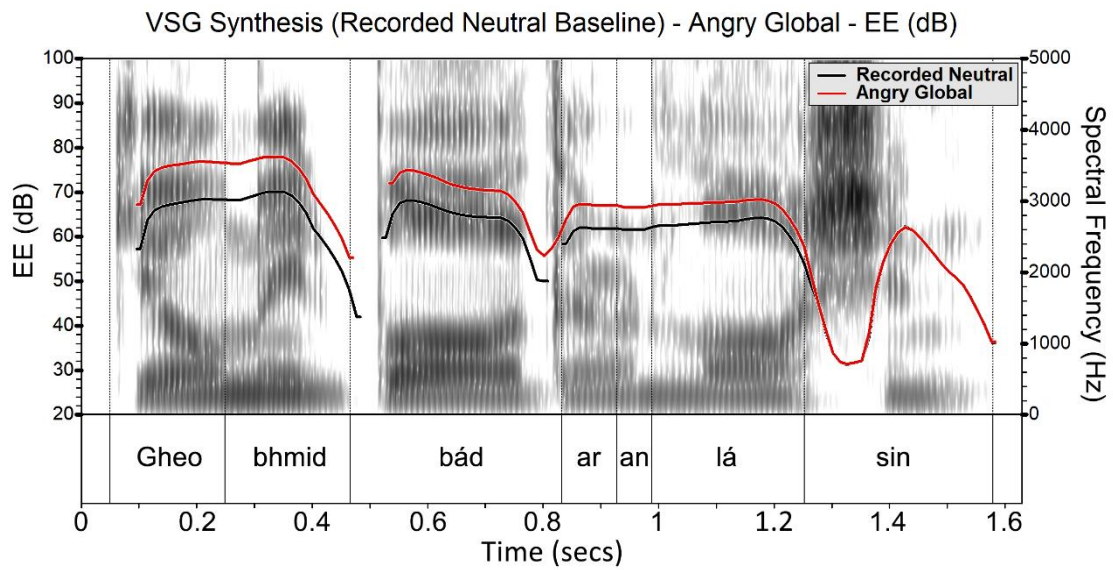


Figure C5 – EE contour of the global modification of the recorded neutral stimulus to the *angry* target affect.

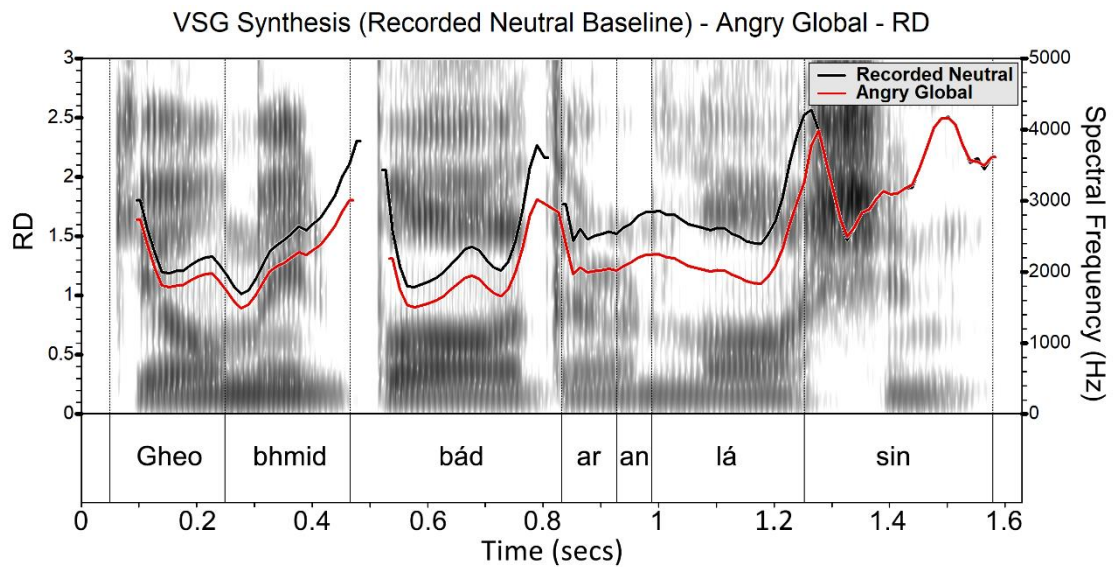


Figure C6 – RD contour of the global modification of the recorded neutral stimulus to the *angry* target affect.

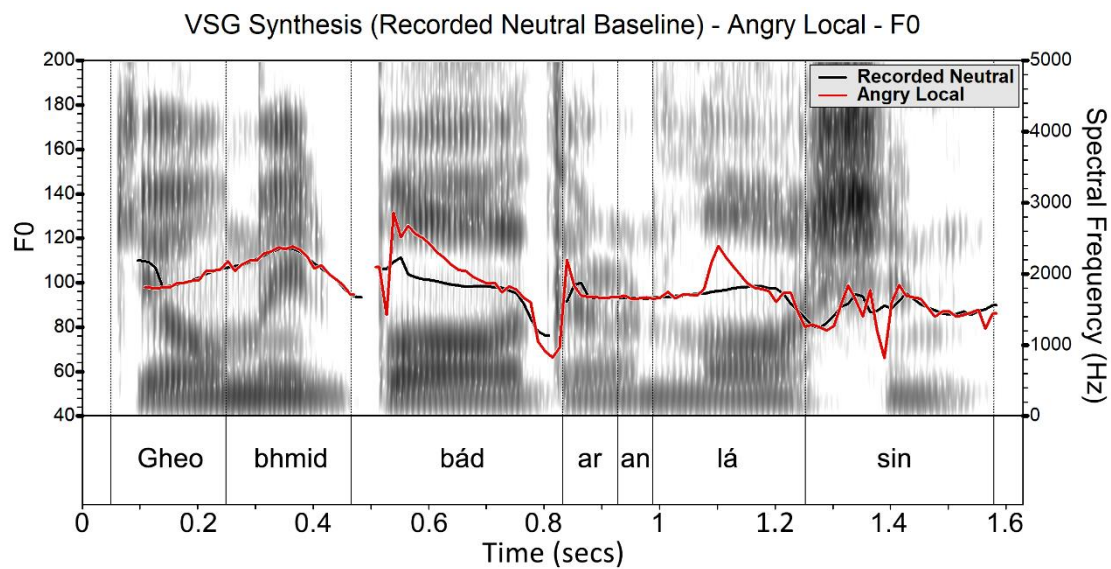


Figure C7 - F_0 contour of the local modification of the recorded neutral stimulus to the *angry* target affect.

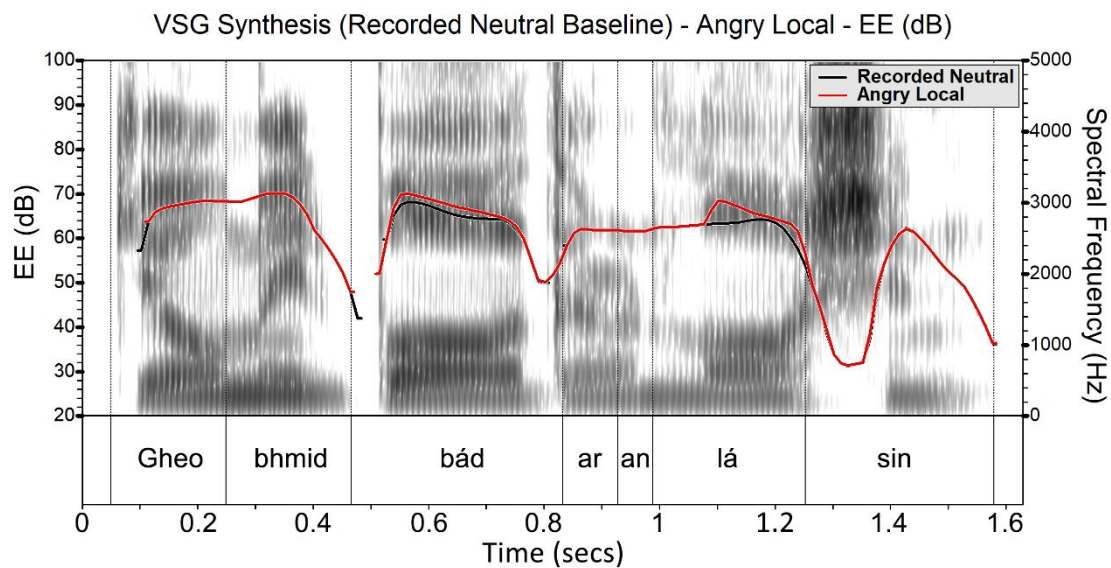


Figure C8 – EE contour of the local modification of the recorded neutral stimulus to the *angry* target affect.

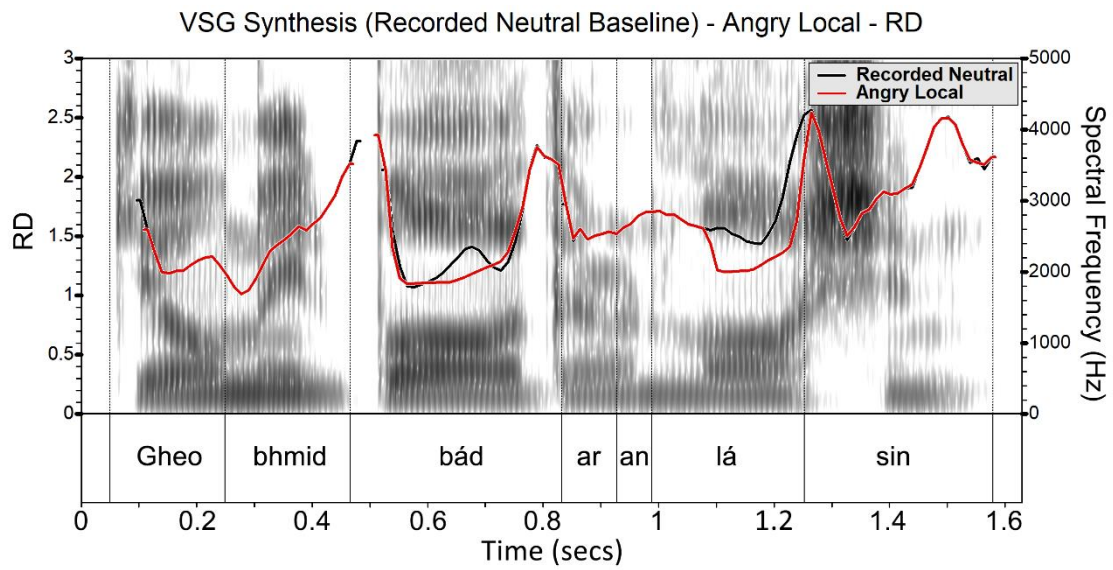


Figure C9 – RD contour of the local modification of the recorded neutral stimulus to the *angry* target affect.

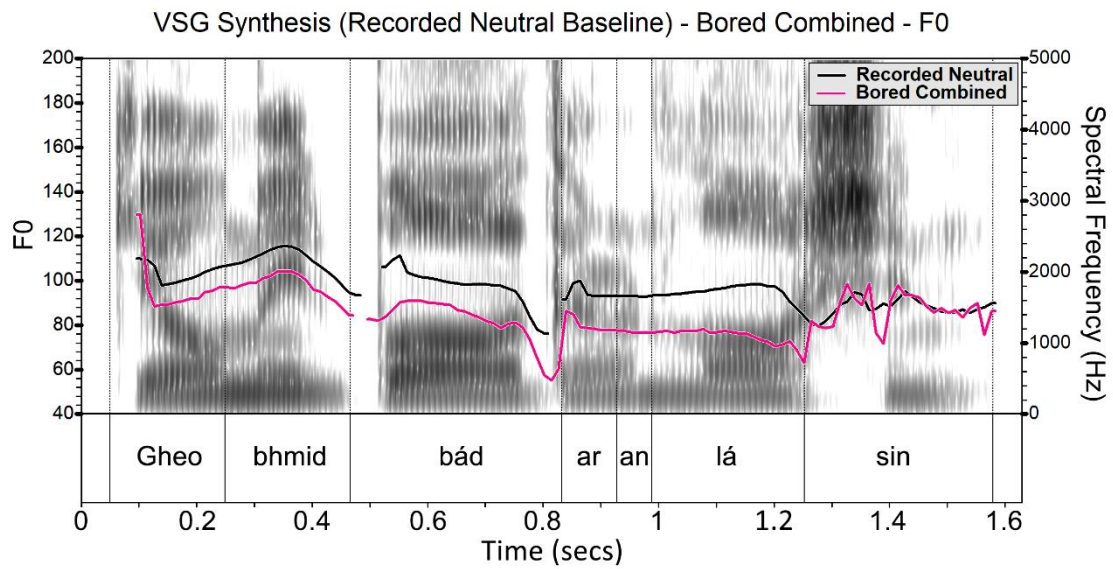


Figure C10 - F_0 contour of the combined modification of the recorded neutral stimulus to the *bored* target affect.

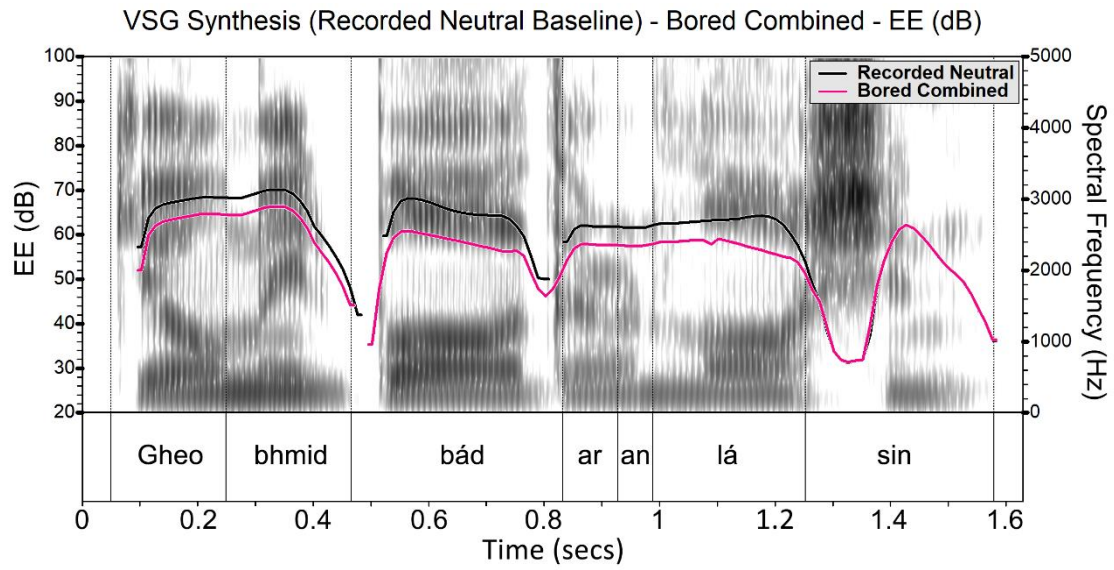


Figure C11 - EE contour of the combined modification of the recorded neutral stimulus to the *bored* target affect.

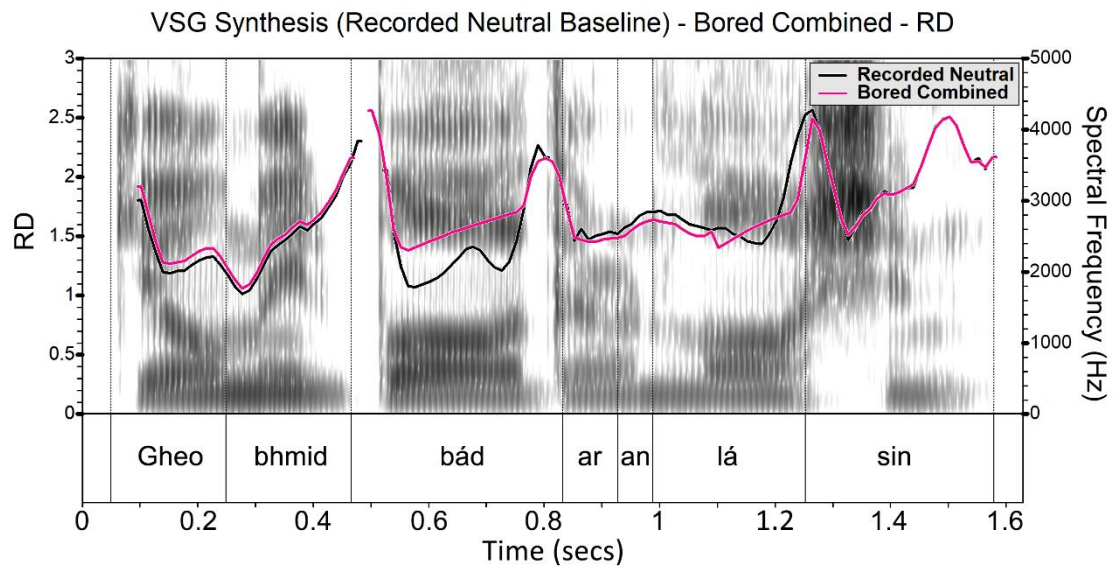


Figure C12 - RD contour of the combined modification of the recorded neutral stimulus to the *bored* target affect.

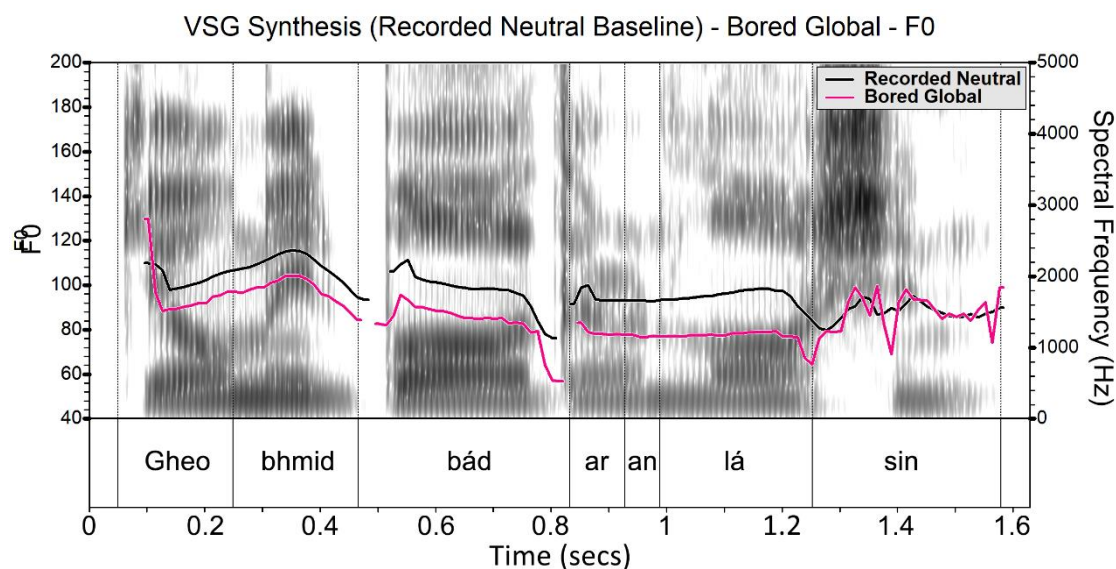


Figure C13 - F_0 contour of the global modification of the recorded neutral stimulus to the *bored* target affect.

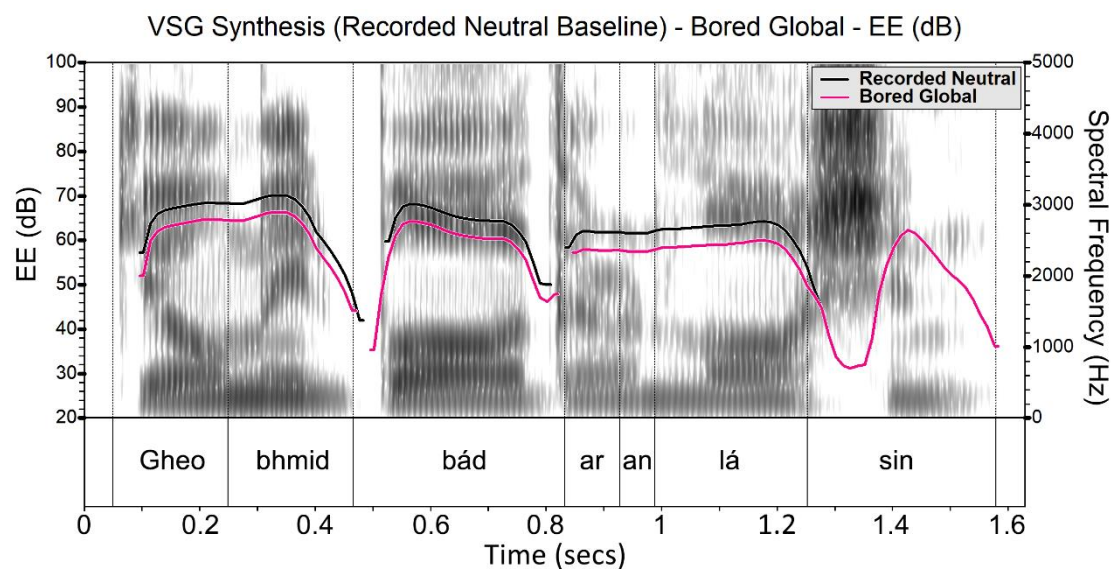


Figure C14 - EE contour of the global modification of the recorded neutral stimulus to the *bored* target affect.

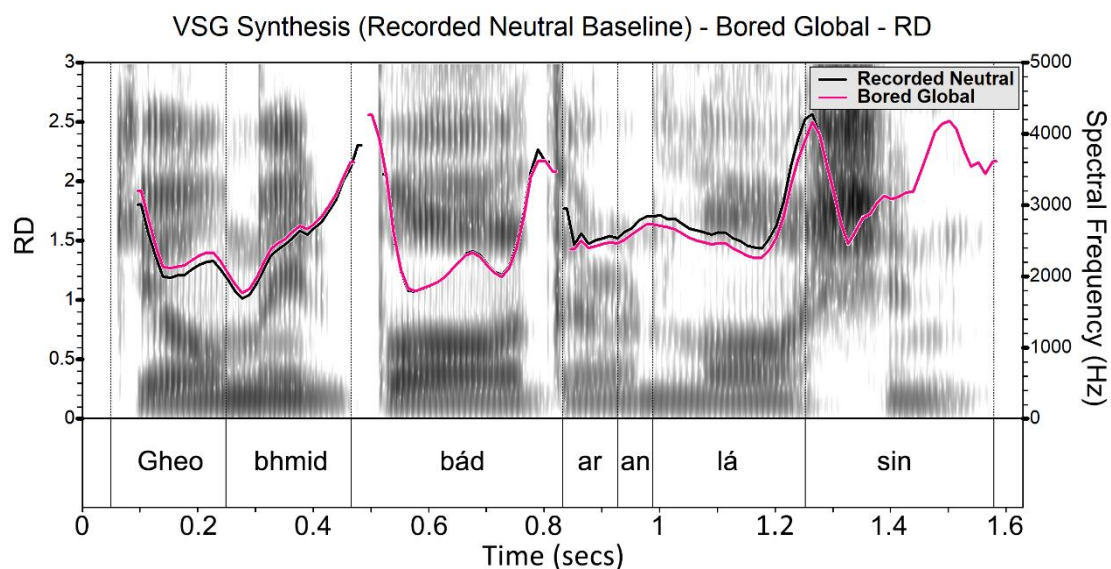


Figure C15 - RD contour of the global modification of the recorded neutral stimulus to the *bored* target affect.

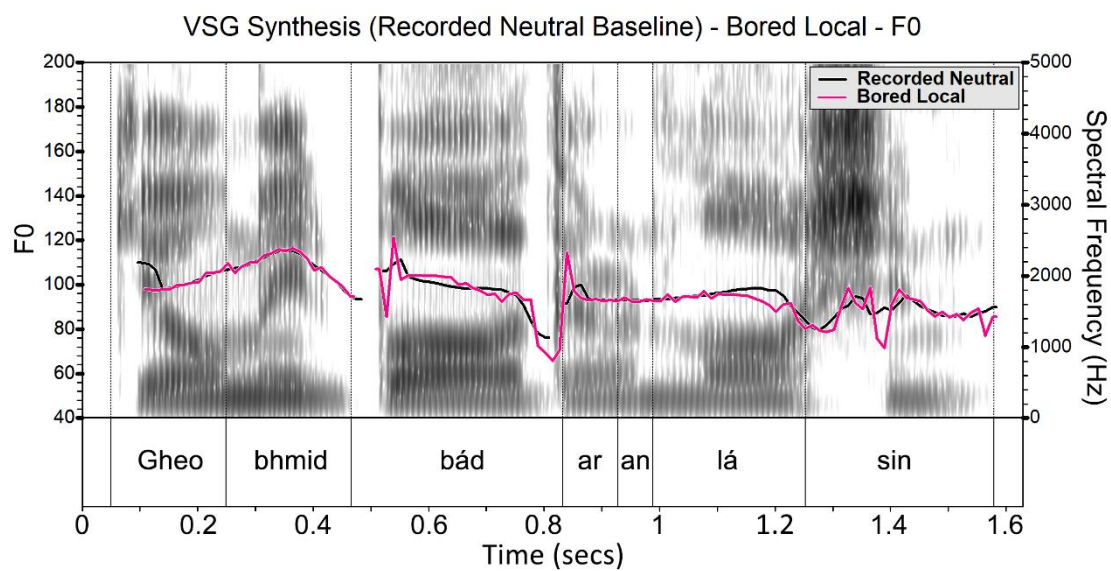


Figure C16 - F_0 contour of the local modification of the recorded neutral stimulus to the *bored* target affect.

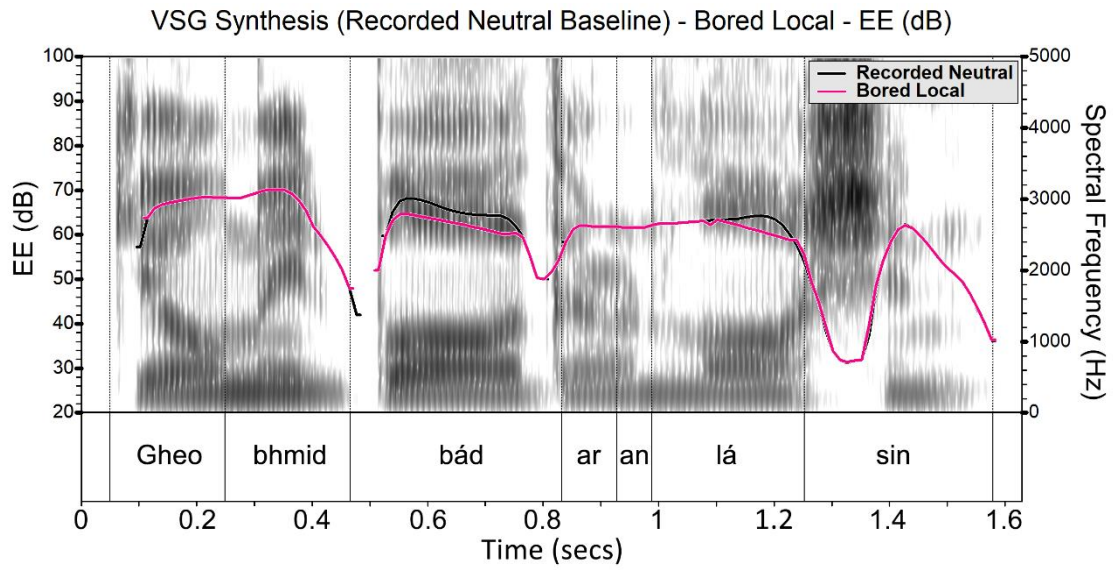


Figure C17 - EE contour of the local modification of the recorded neutral stimulus to the *bored* target affect.

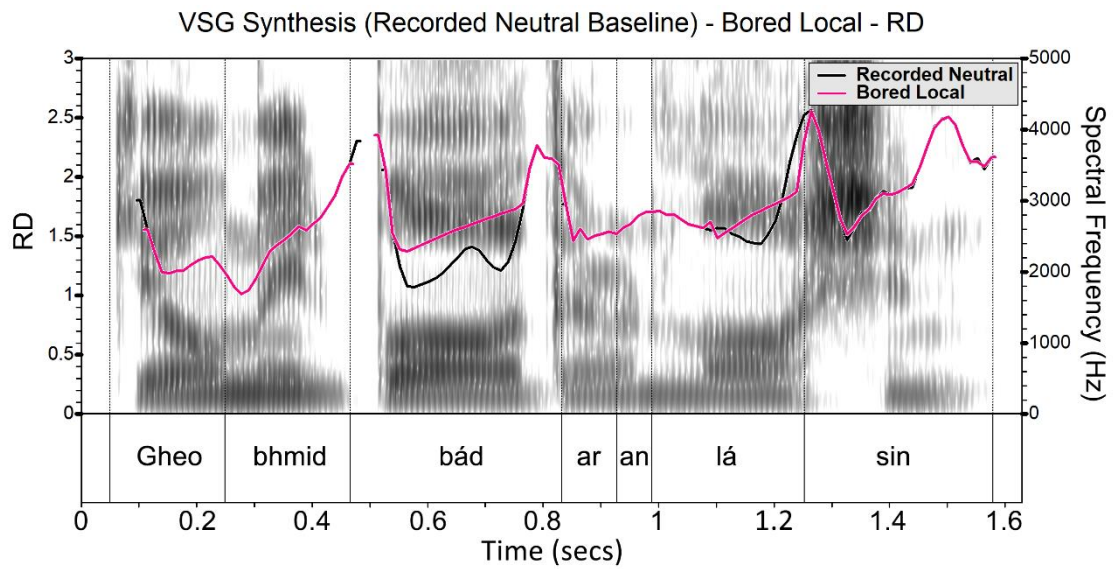


Figure C18 - RD contour of the local modification of the recorded neutral stimulus to the *bored* target affect.

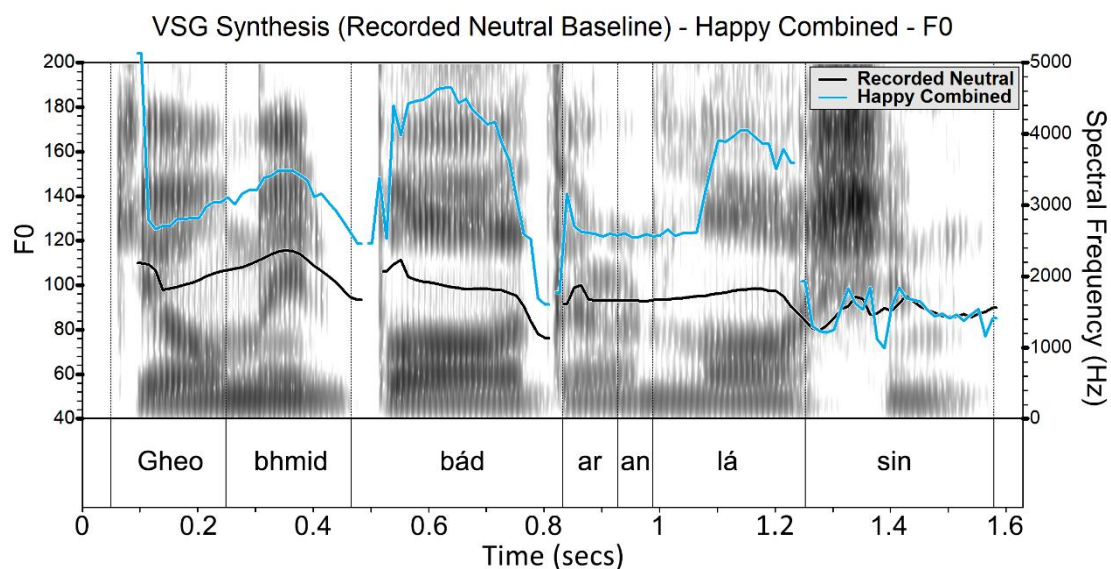


Figure C19 - F_0 contour of the combined modification of the recorded neutral stimulus to the *happy* target affect.

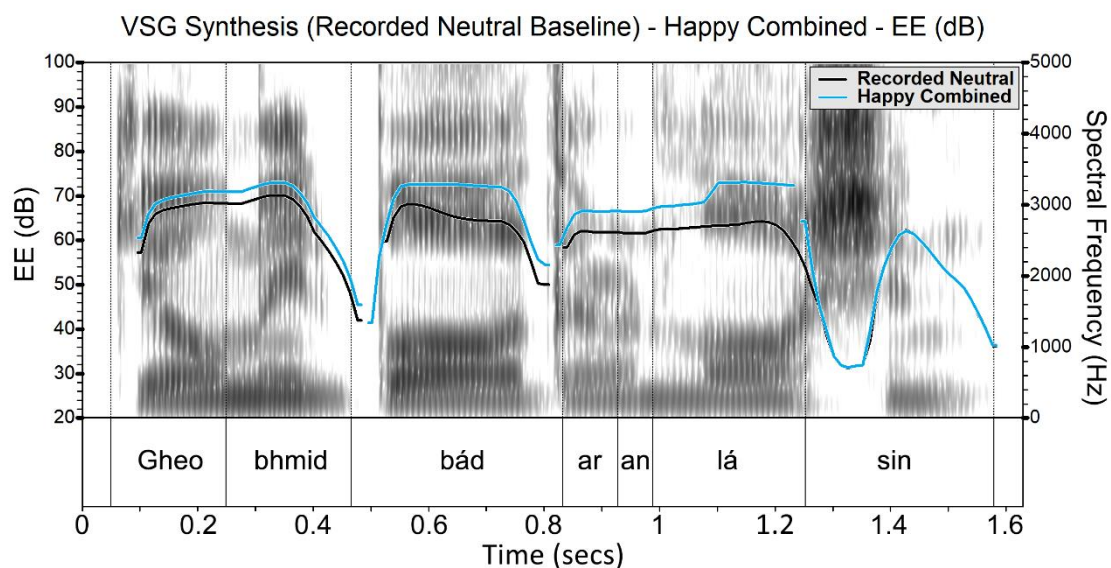


Figure C20 - EE contour of the combined modification of the recorded neutral stimulus to the *happy* target affect.

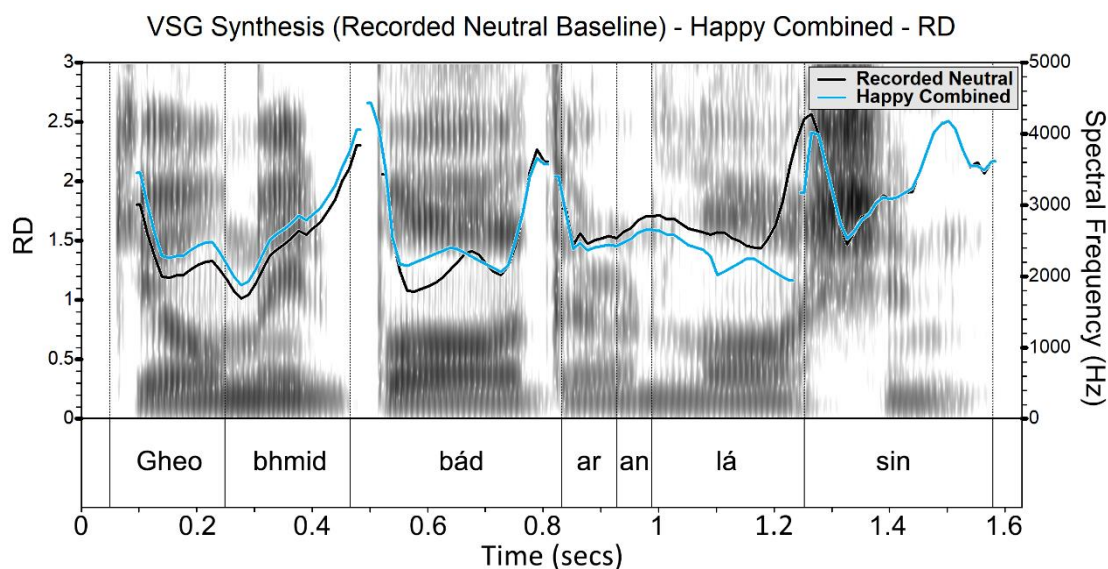


Figure C21 - RD contour of the combined modification of the recorded neutral stimulus to the *happy* target affect.

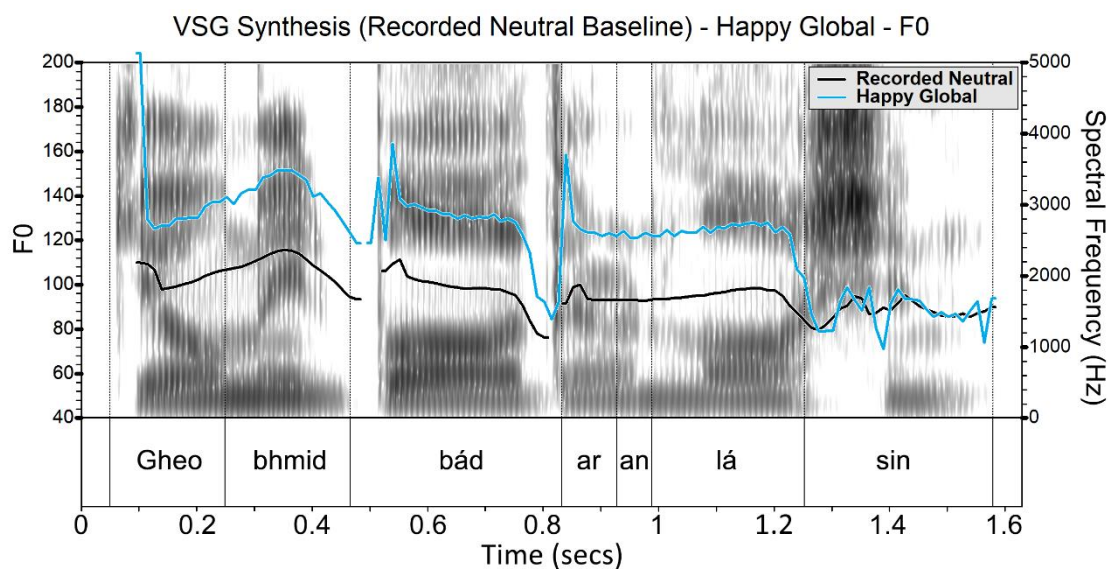


Figure C22 - F₀ contour of the global modification of the recorded neutral stimulus to the *happy* target affect.

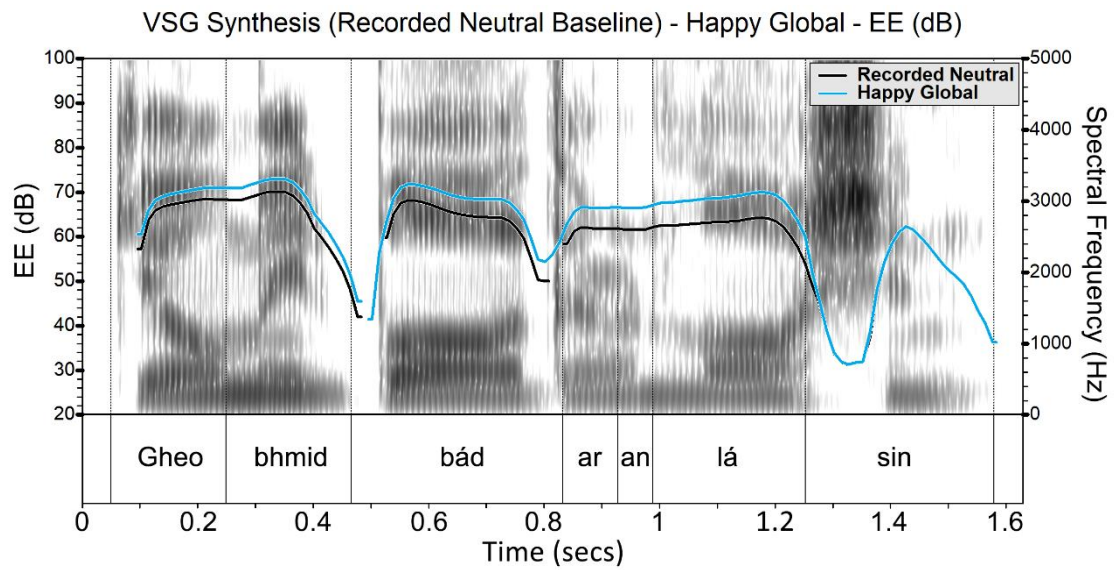


Figure C23 - EE contour of the combined modification of the recorded neutral stimulus to the *happy* target affect.

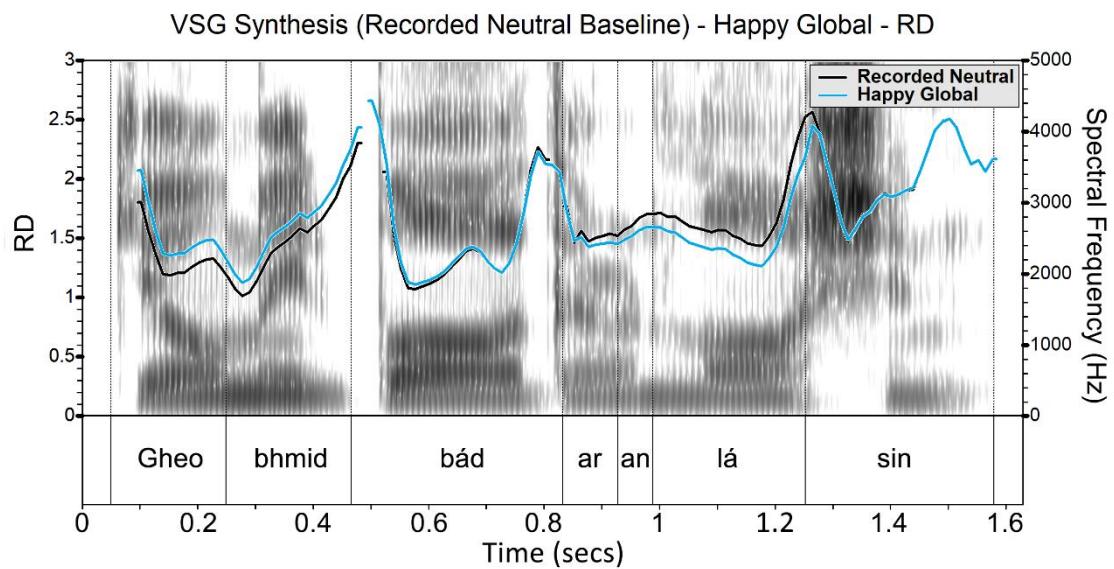


Figure C24 - RD contour of the global modification of the recorded neutral stimulus to the *happy* target affect.

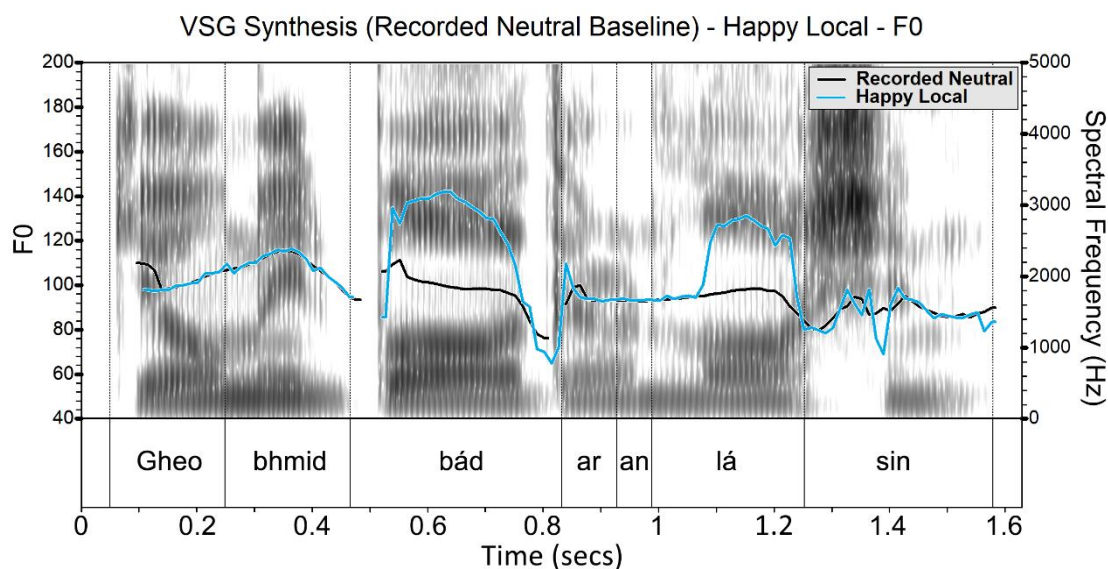


Figure C25 - F_0 contour of the local modification of the recorded neutral stimulus to the *happy* target affect.

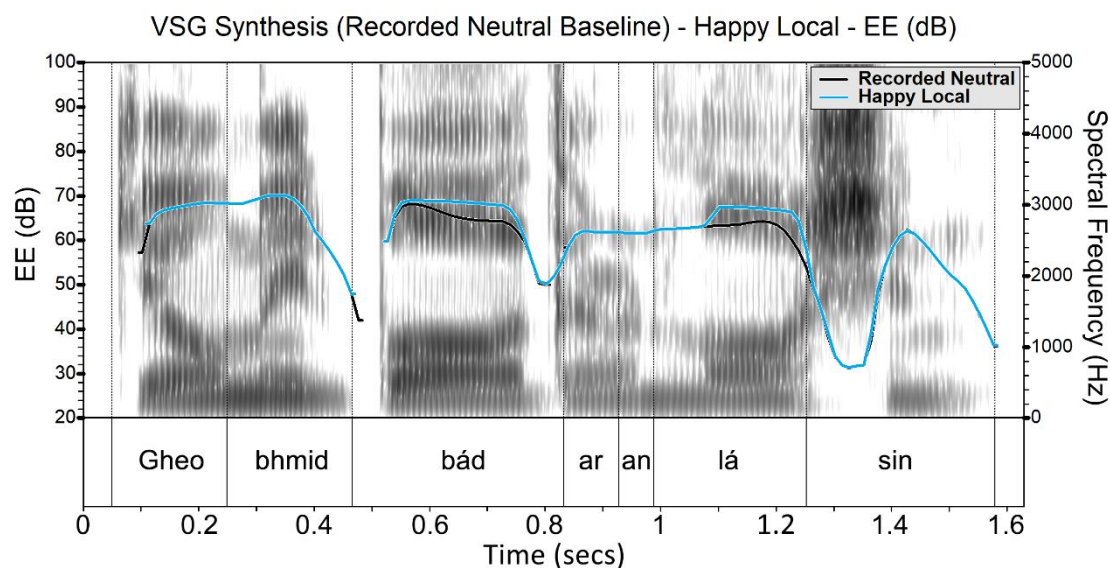


Figure C26 - EE contour of the combined modification of the recorded neutral stimulus to the *happy* target affect.

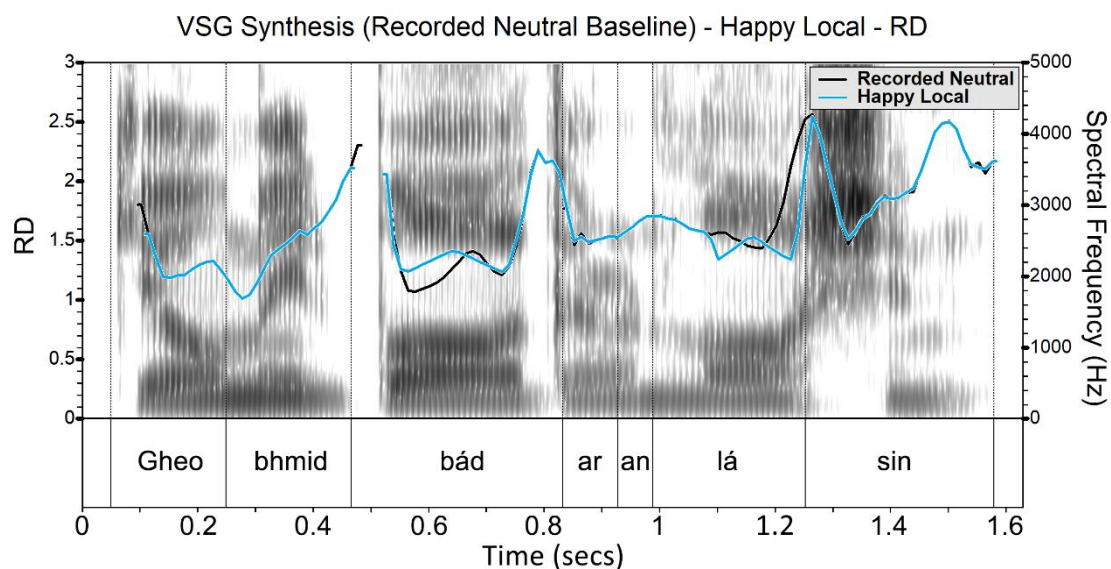


Figure C27 – RD contour of the combined modification of the recorded neutral stimulus to the *happy* target affect.

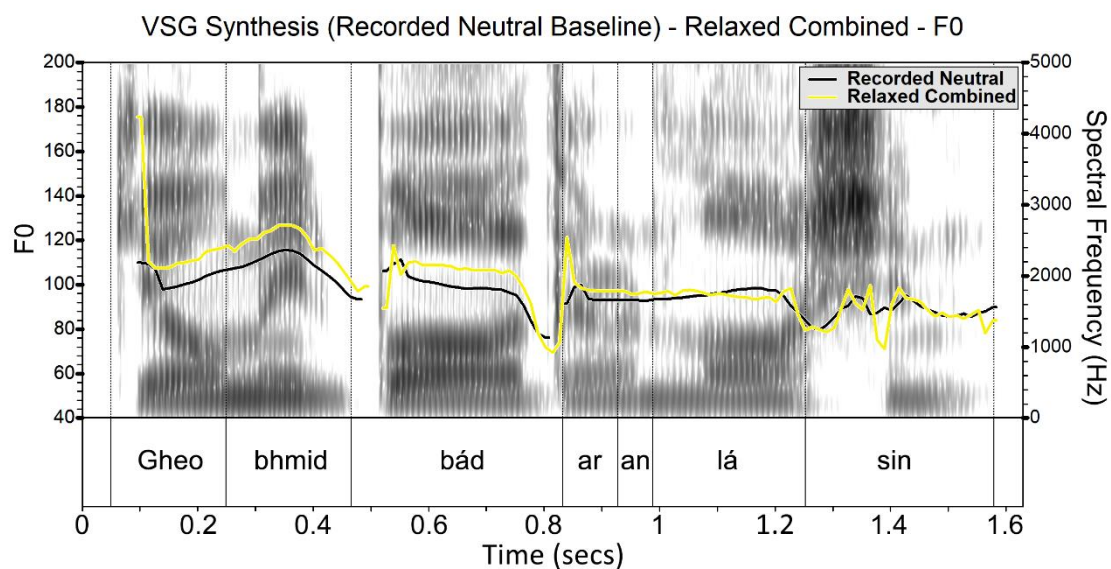


Figure C28 - F_0 contour of the combined modification of the recorded neutral stimulus to the *relaxed* target affect.

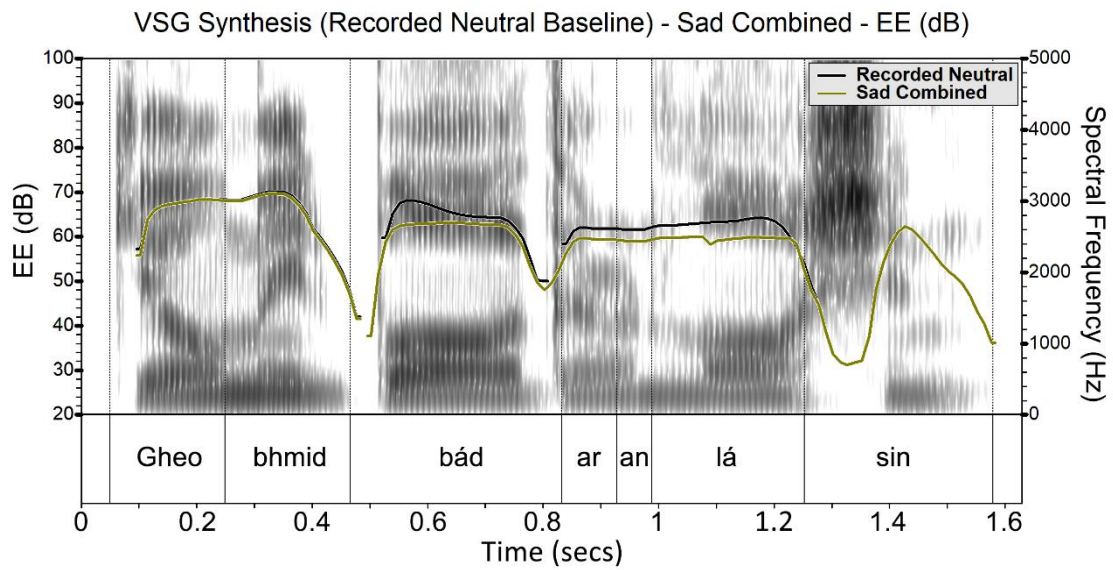


Figure C29 - EE contour of the combined modification of the recorded neutral stimulus to the *relaxed* target affect.

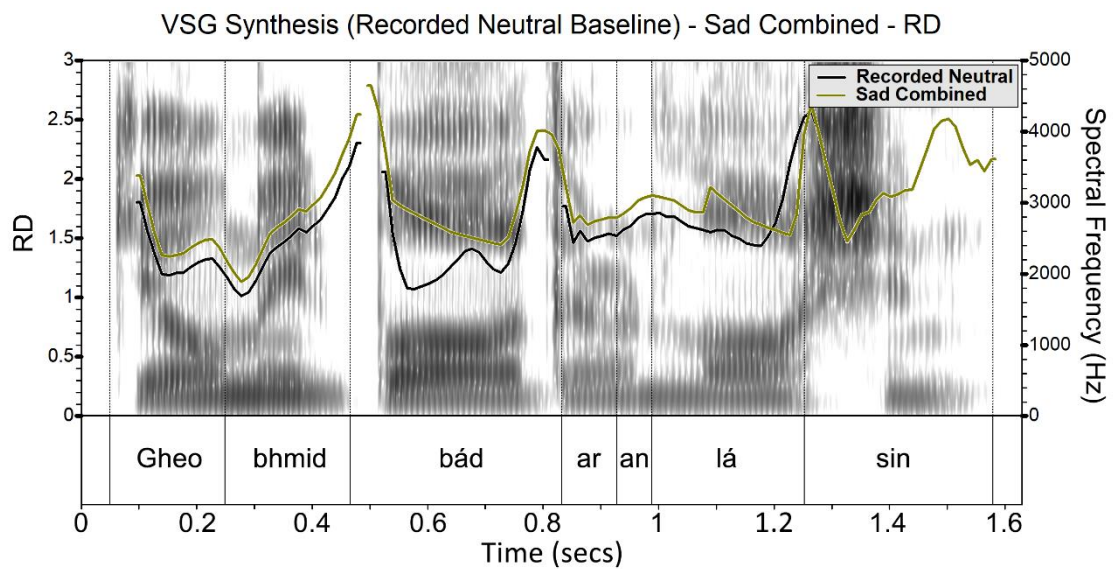


Figure C30 - RD contour of the combined modification of the recorded neutral stimulus to the *relaxed* target affect.

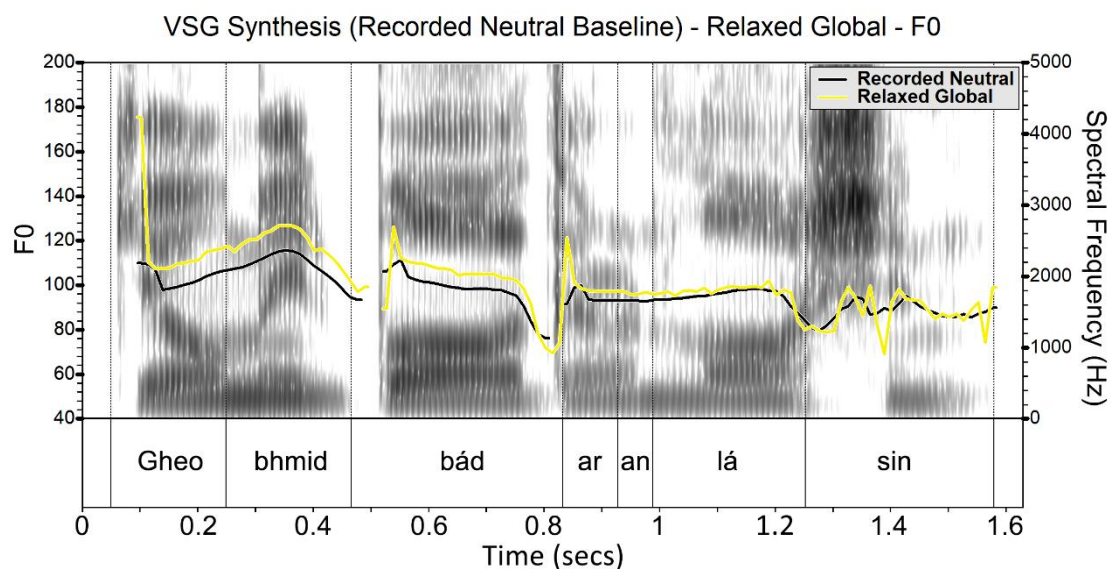


Figure C31 - F_0 contour of the global modification of the recorded neutral stimulus to the *relaxed* target affect.

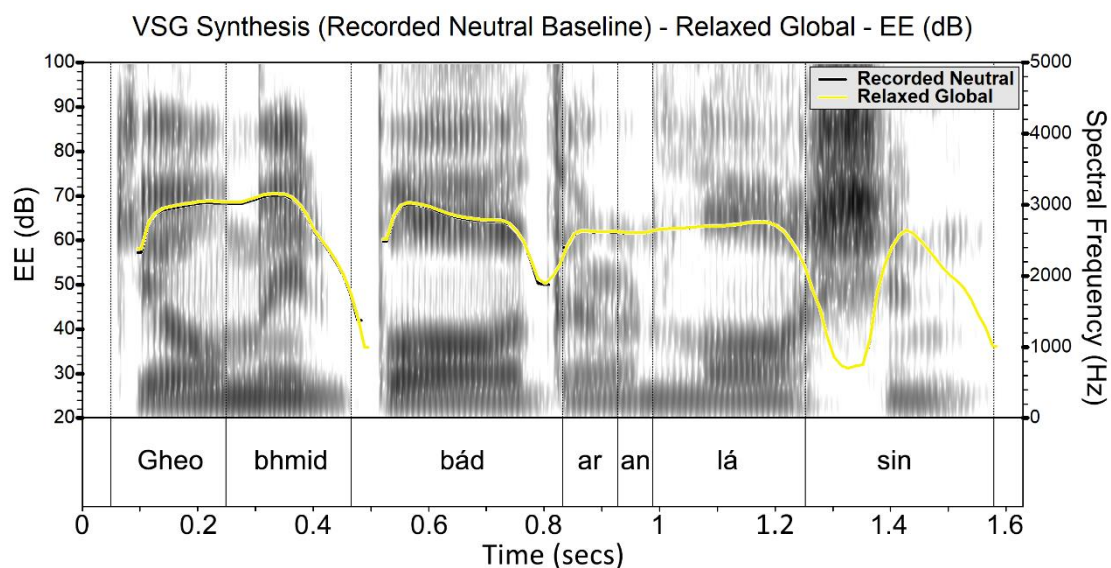


Figure C32 - EE contour of the global modification of the recorded neutral stimulus to the *relaxed* target affect.

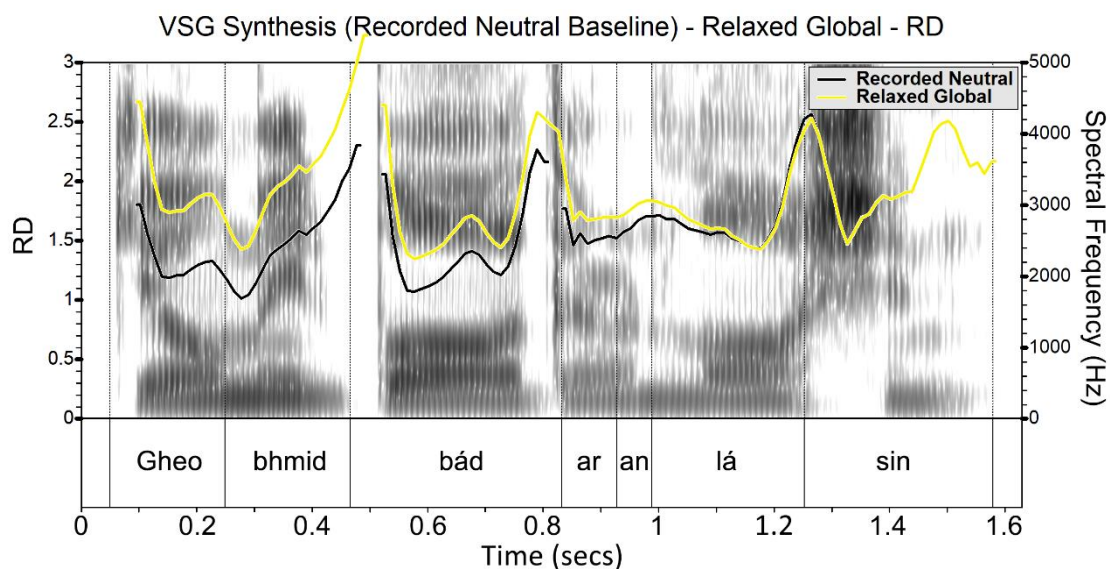


Figure C33 - RD contour of the global modification of the recorded neutral stimulus to the *relaxed* target affect.

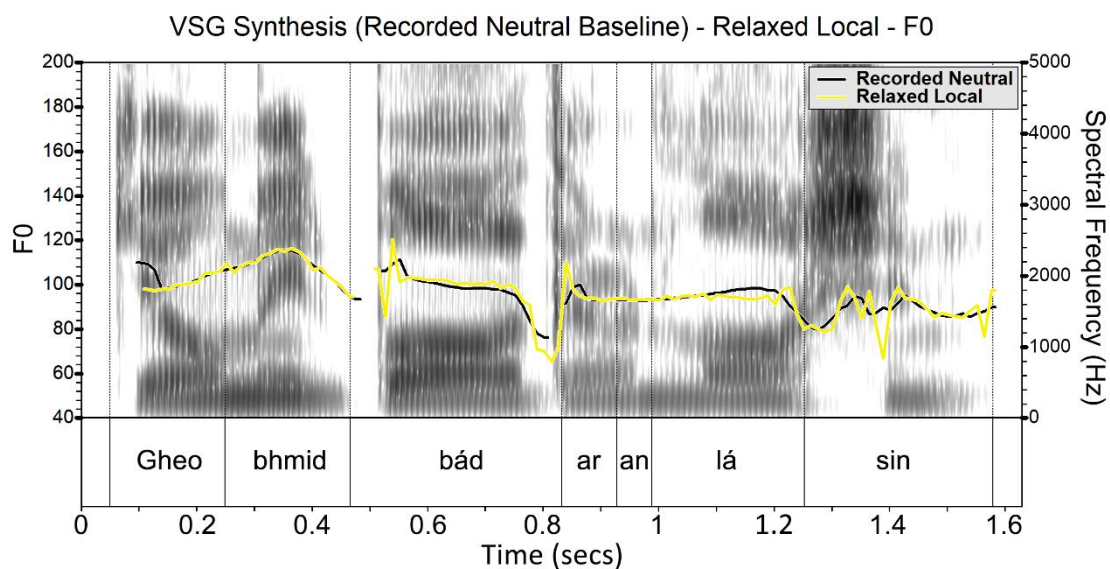


Figure C34 - F₀ contour of the local modification of the recorded neutral stimulus to the *relaxed* target affect.

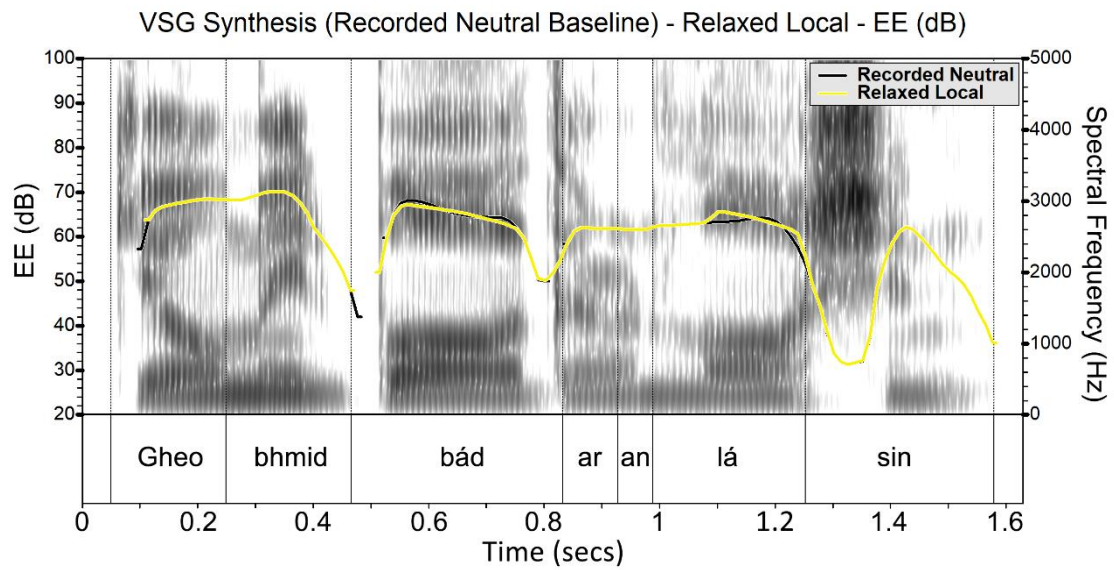


Figure C35 - EE contour of the local modification of the recorded neutral stimulus to the *relaxed* target affect.

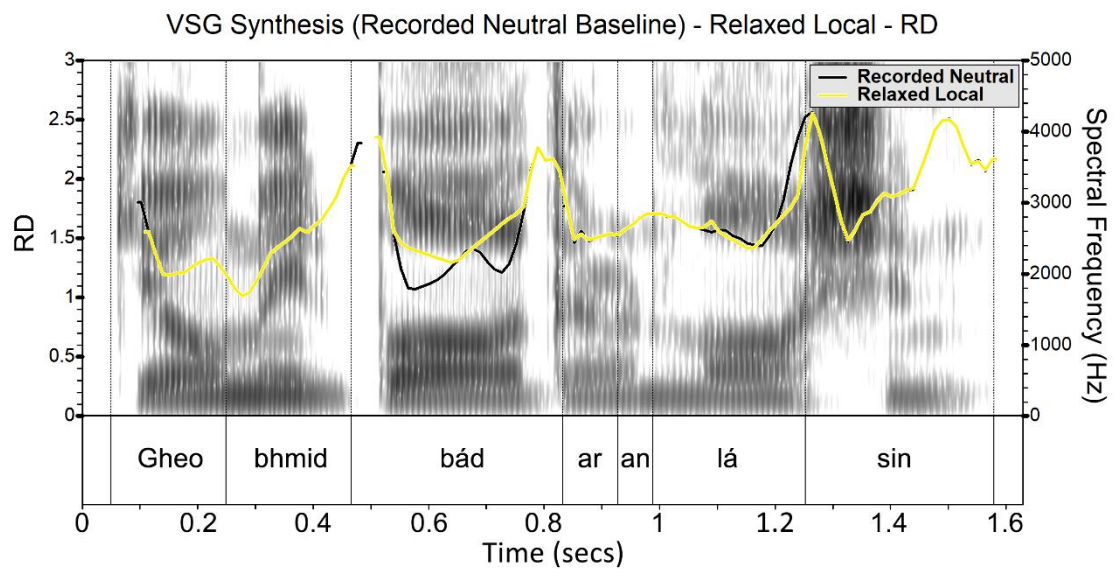


Figure C36 - RD contour of the local modification of the recorded neutral stimulus to the *relaxed* target affect.

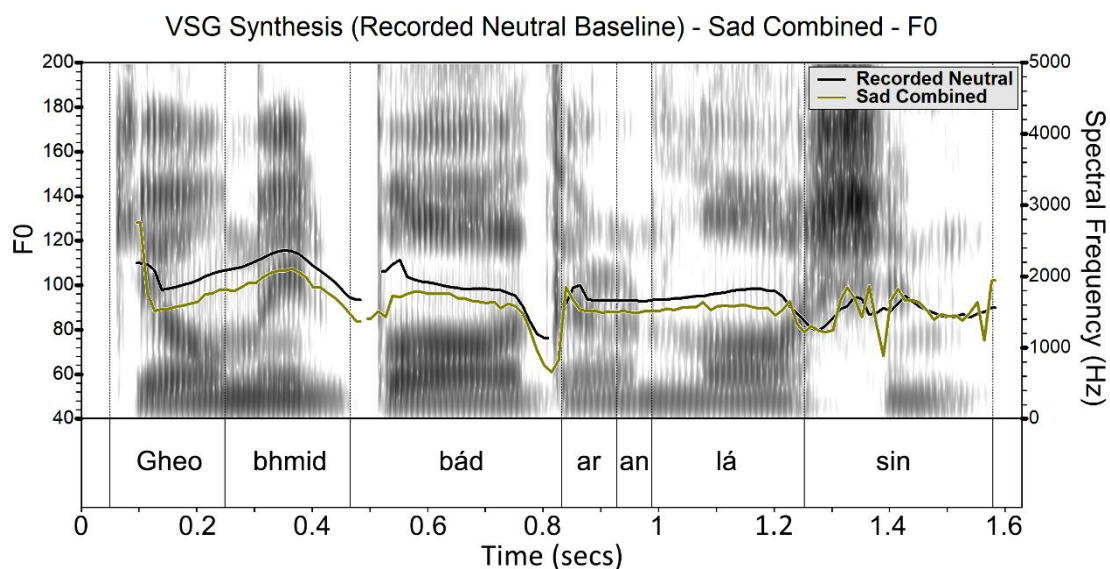


Figure C37 - F_0 contour of the combined modification of the recorded neutral stimulus to the *sad* target affect.

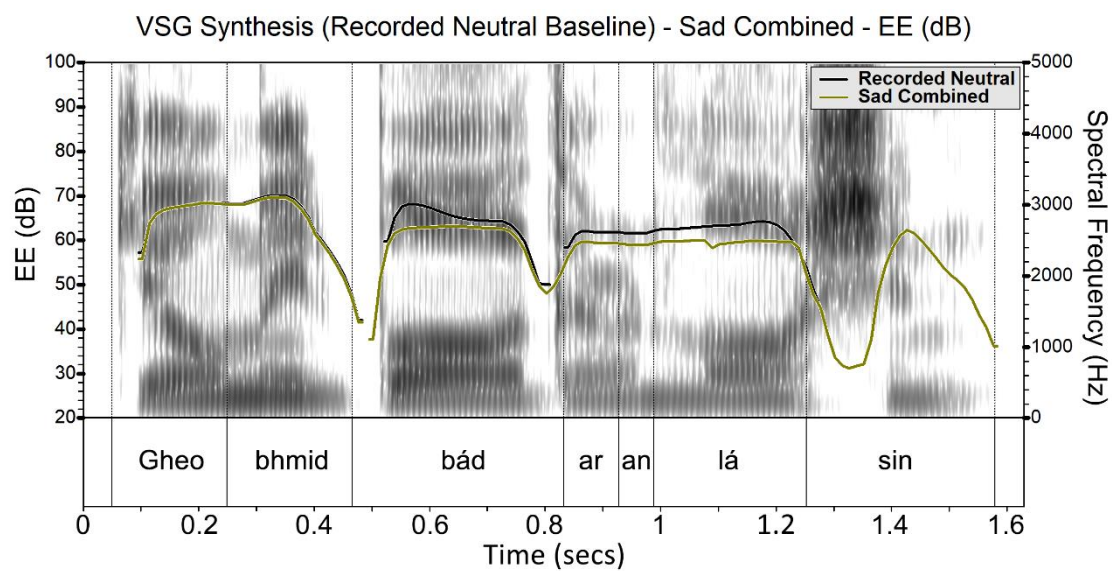


Figure C38 - EE contour of the combined modification of the recorded neutral stimulus to the *sad* target affect.

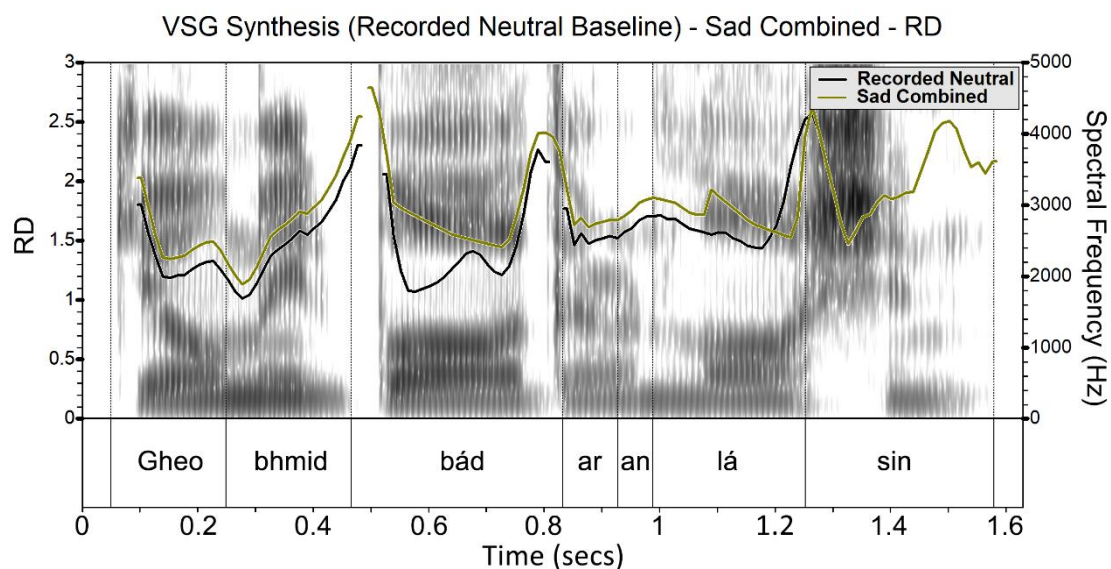


Figure C39 - RD contour of the combined modification of the recorded neutral stimulus to the *sad* target affect.

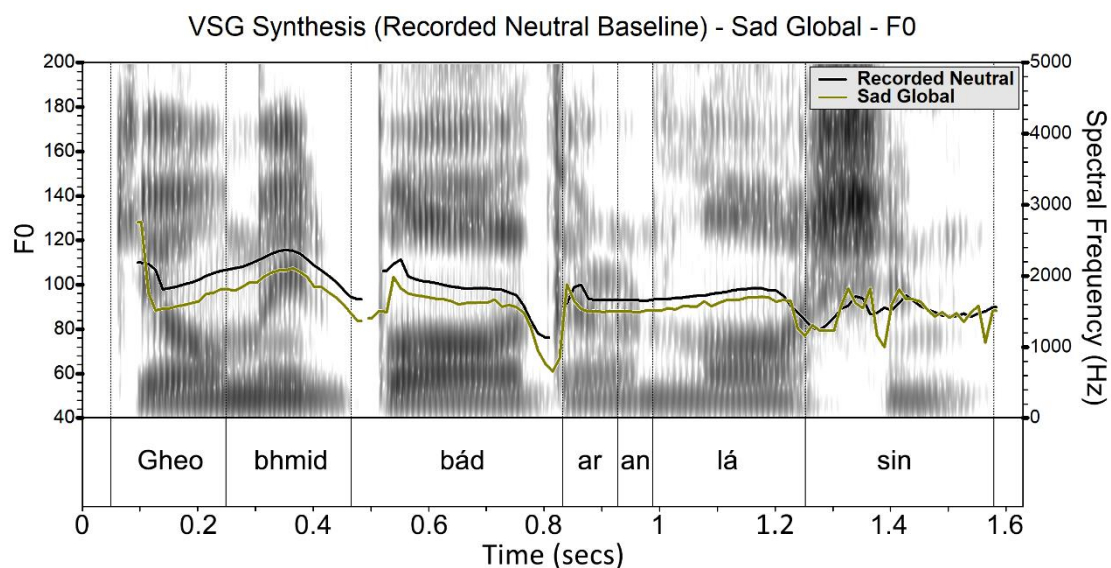


Figure C40 - F_0 contour of the global modification of the recorded neutral stimulus to the *sad* target affect.

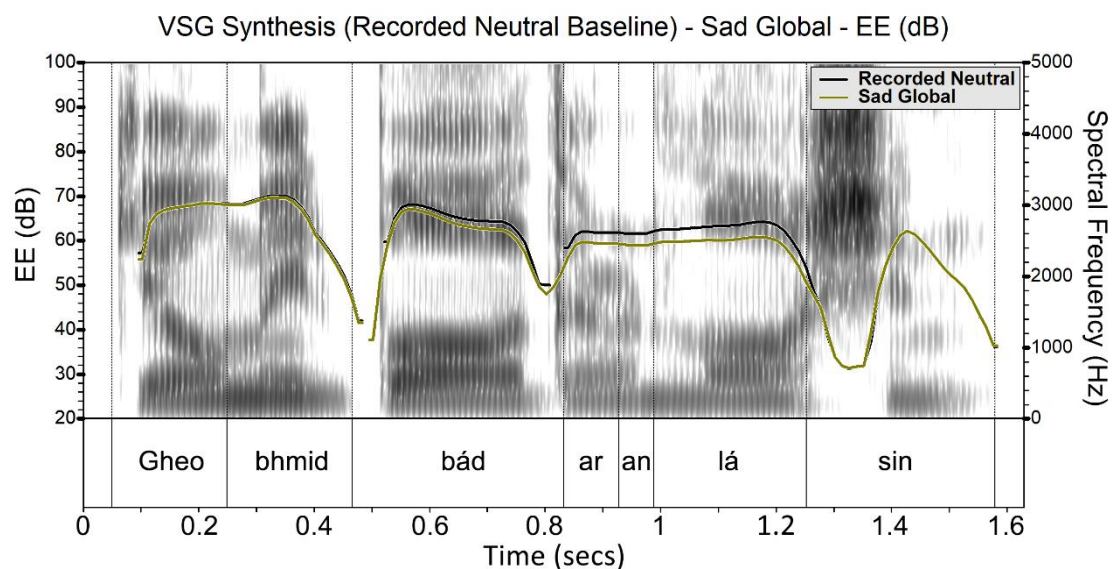


Figure C41 - EE contour of the global modification of the recorded neutral stimulus to the *sad* target affect.

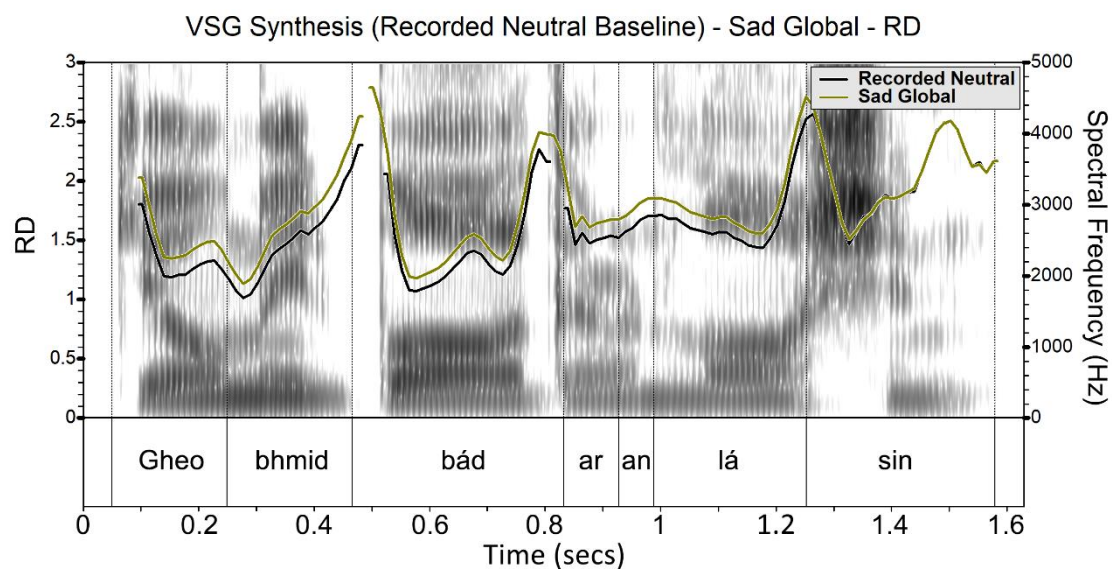


Figure C42 - RD contour of the global modification of the recorded neutral stimulus to the *sad* target affect.

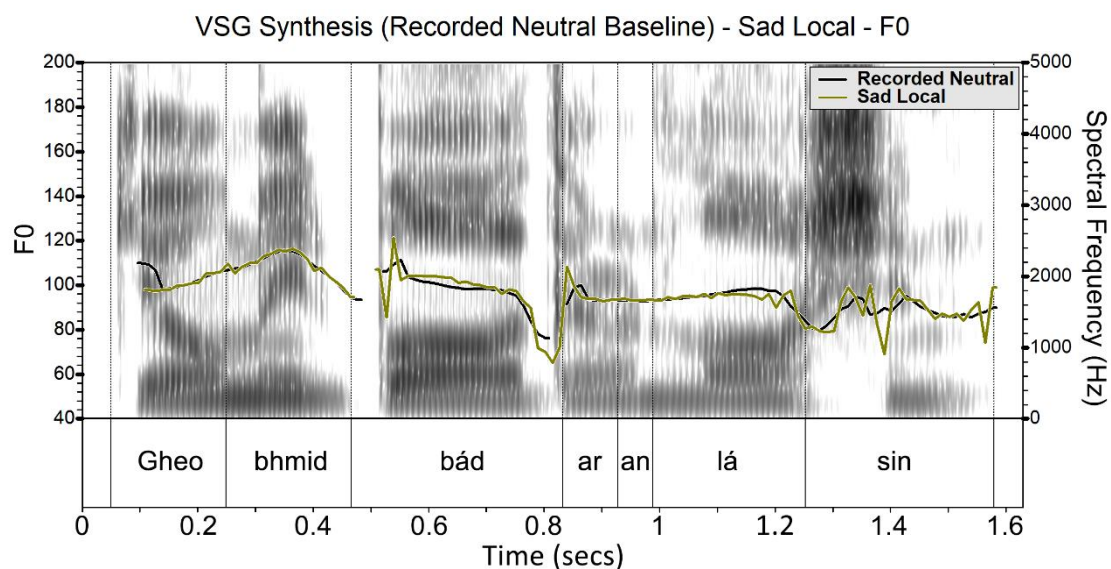


Figure C43 - F_0 contour of the local modification of the recorded neutral stimulus to the *sad* target affect.

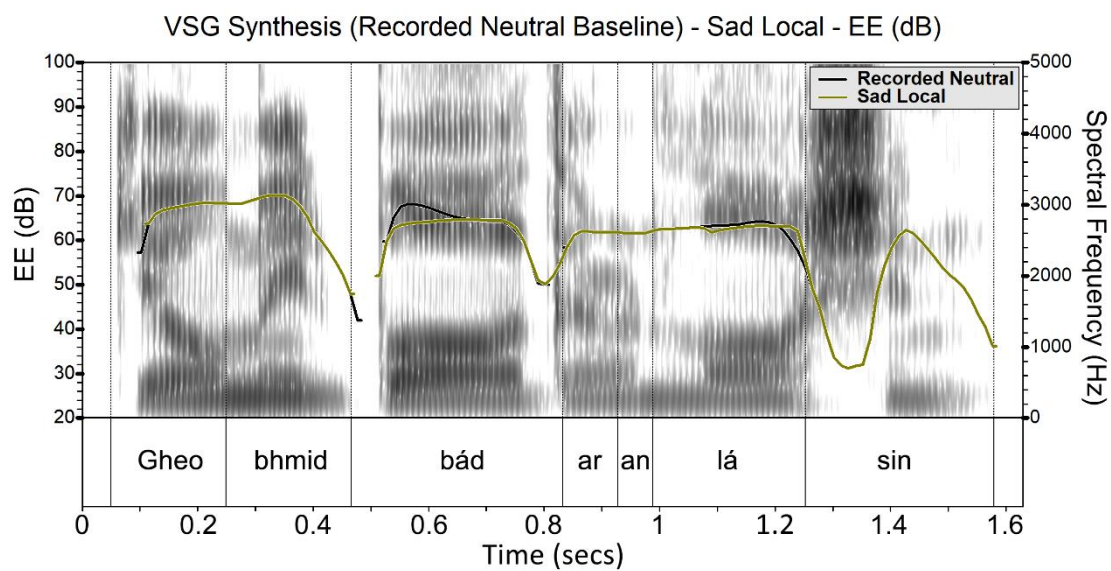


Figure C44 - EE contour of the local modification of the recorded neutral stimulus to the *sad* target affect.

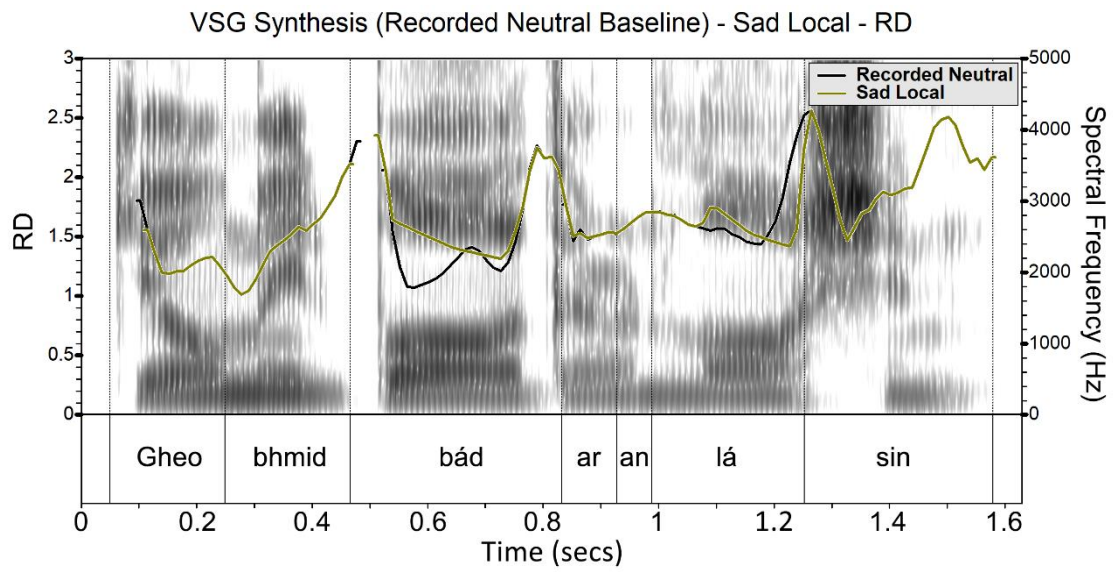


Figure C45 - RD contour of the local modification of the recorded neutral stimulus to the *sad* target affect.

Appendix D

All voice source parameter shifts for all affects for the ABAIR neutral baseline experiment and the visual context experiments.

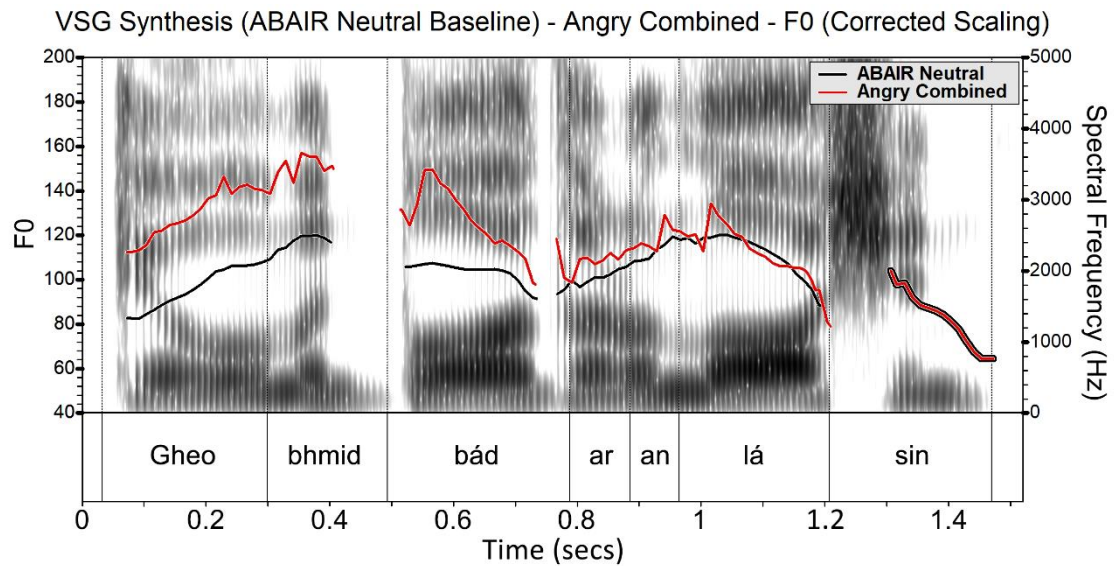


Figure D1 - F₀ contour of the combined modification of the rescaled ABAIR *neutral* stimulus to the *angry* target affect.

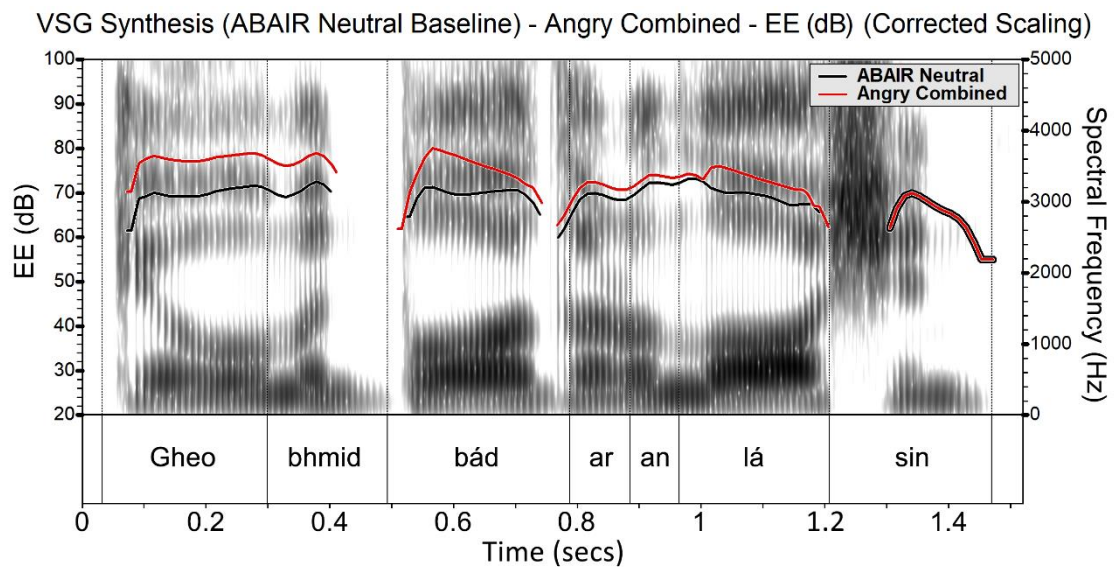


Figure D2 - EE contour of the combined modification of the rescaled ABAIR *neutral* stimulus to the *angry* target affect.

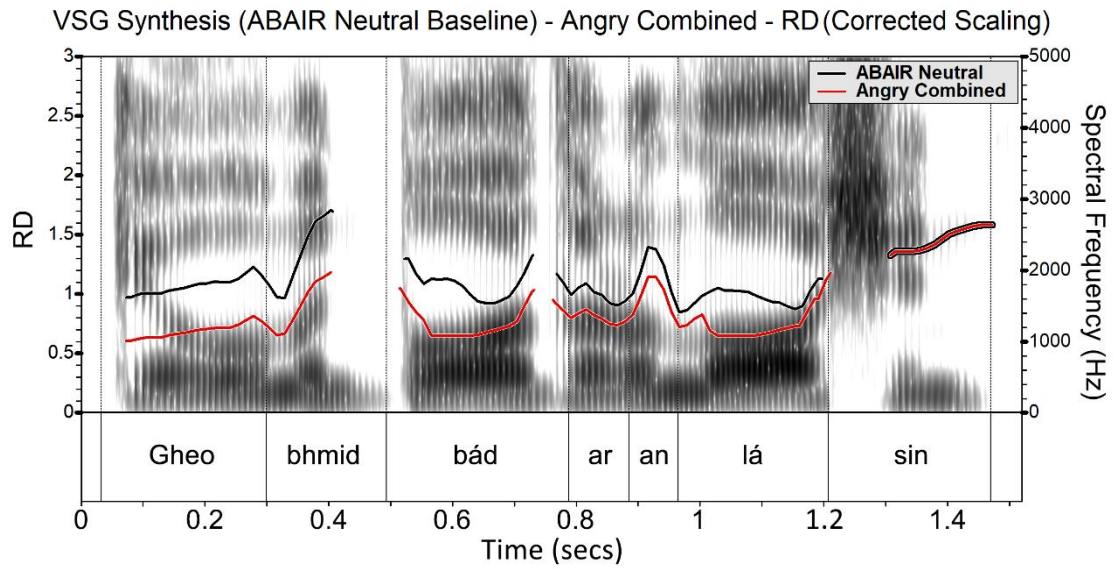


Figure D3 - RD contour of the combined modification of the rescaled ABAIR *neutral* stimulus to the *angry* target affect.

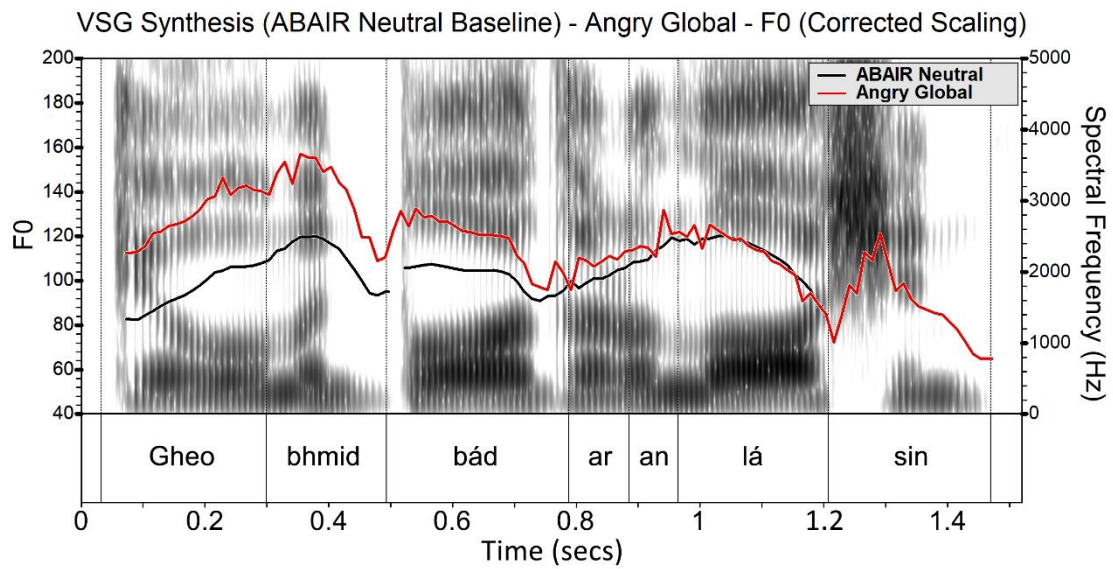


Figure D4 - F_0 contour of the global modification of the rescaled ABAIR *neutral* stimulus to the *angry* target affect.

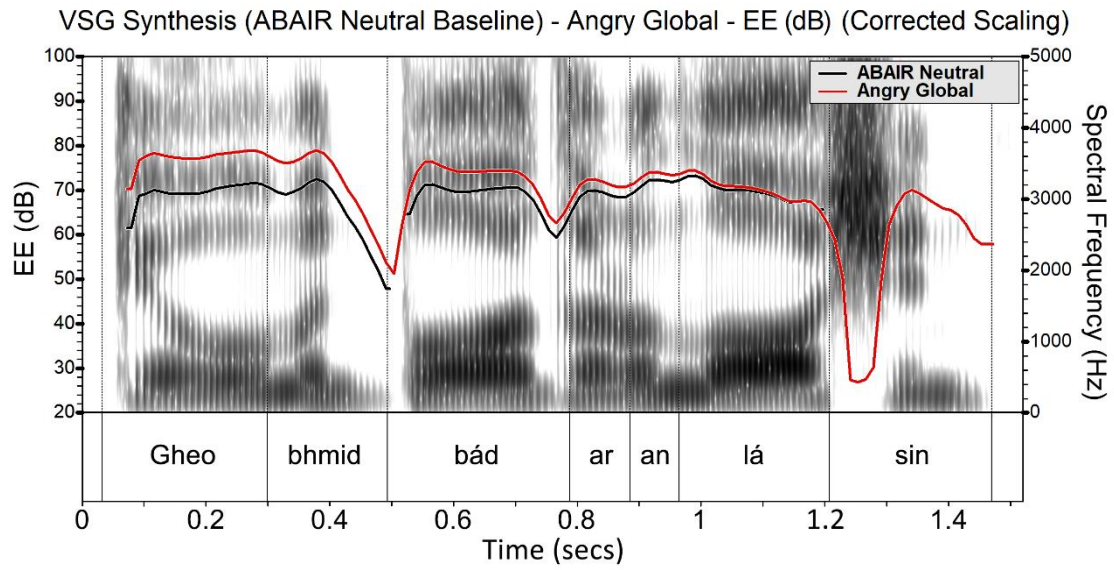


Figure D5 - EE contour of the global modification of the rescaled ABAIR *neutral* stimulus to the *angry* target affect.

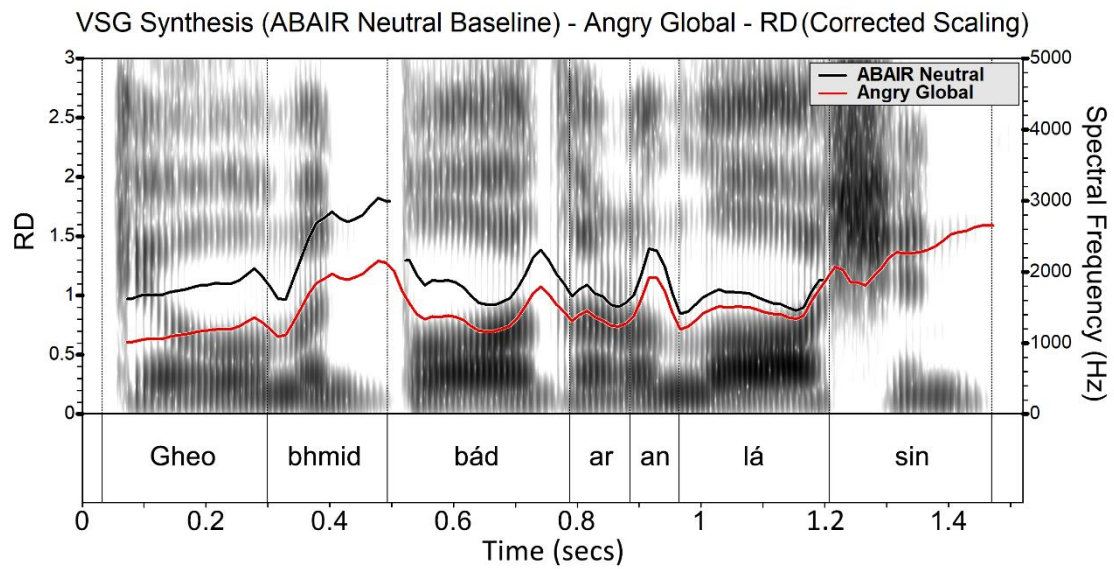


Figure D6 - RD contour of the global modification of the rescaled ABAIR *neutral* stimulus to the *angry* target affect.

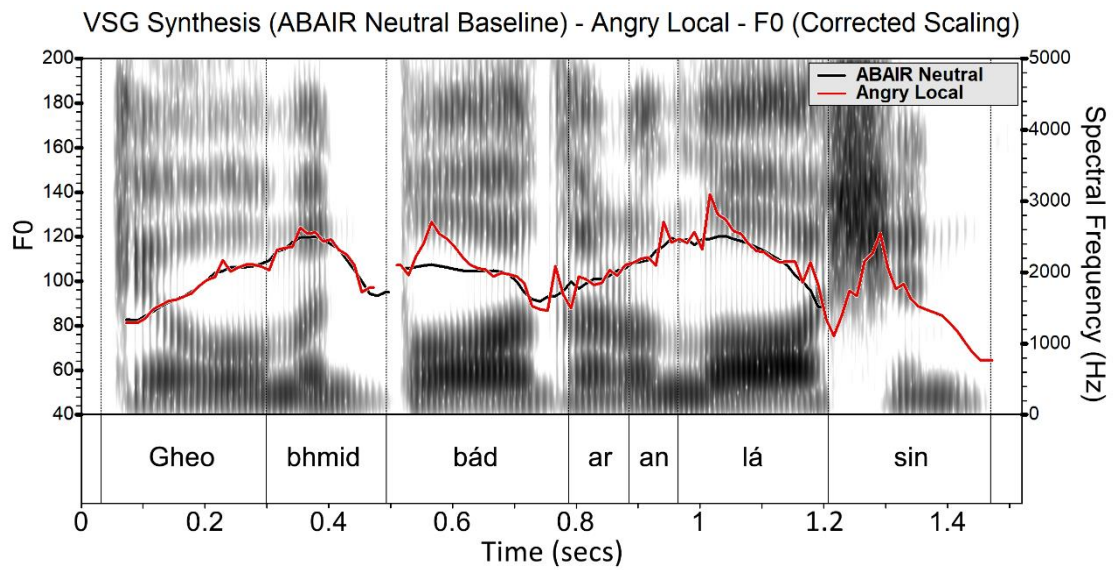


Figure D7 - F₀ contour of the local modification of the rescaled ABAIR *neutral* stimulus to the *angry* target affect.

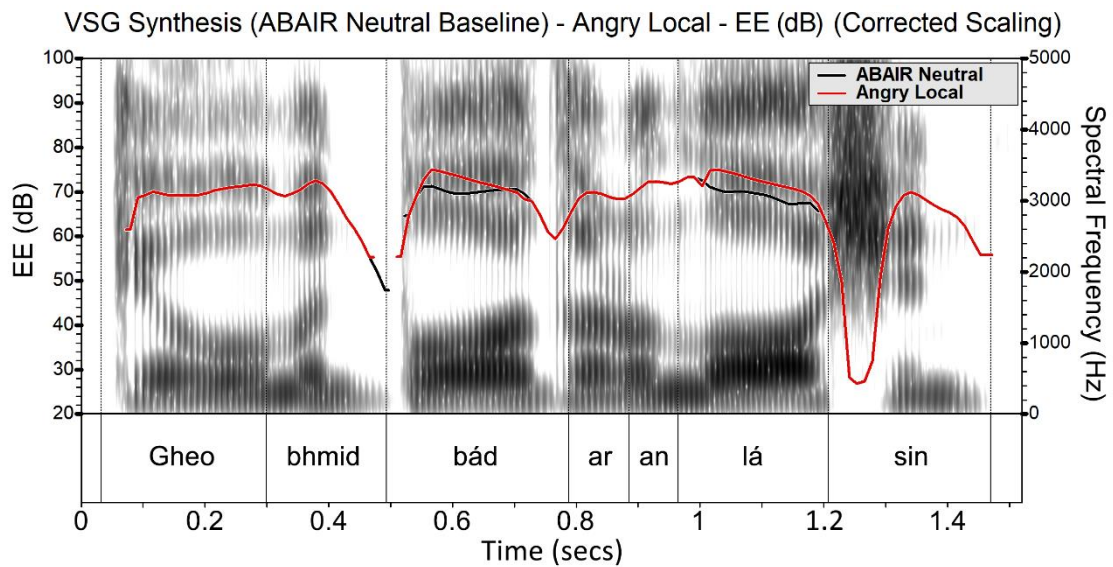


Figure D8 - EE contour of the local modification of the rescaled ABAIR *neutral* stimulus to the *angry* target affect.

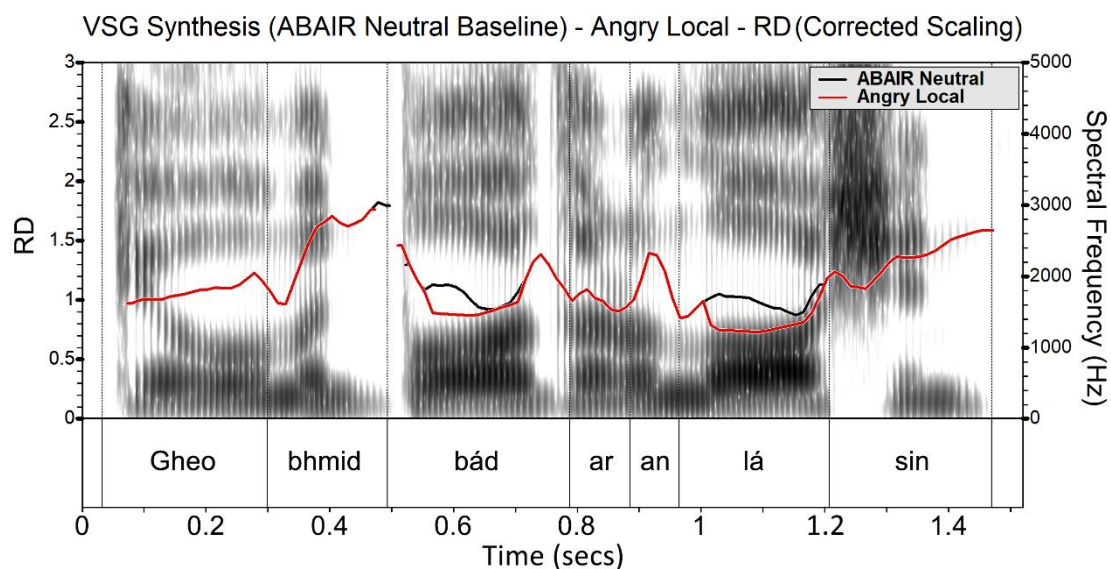


Figure D9 –RD contour of the local modification of the rescaled ABAIR *neutral* stimulus to the *angry* target affect.

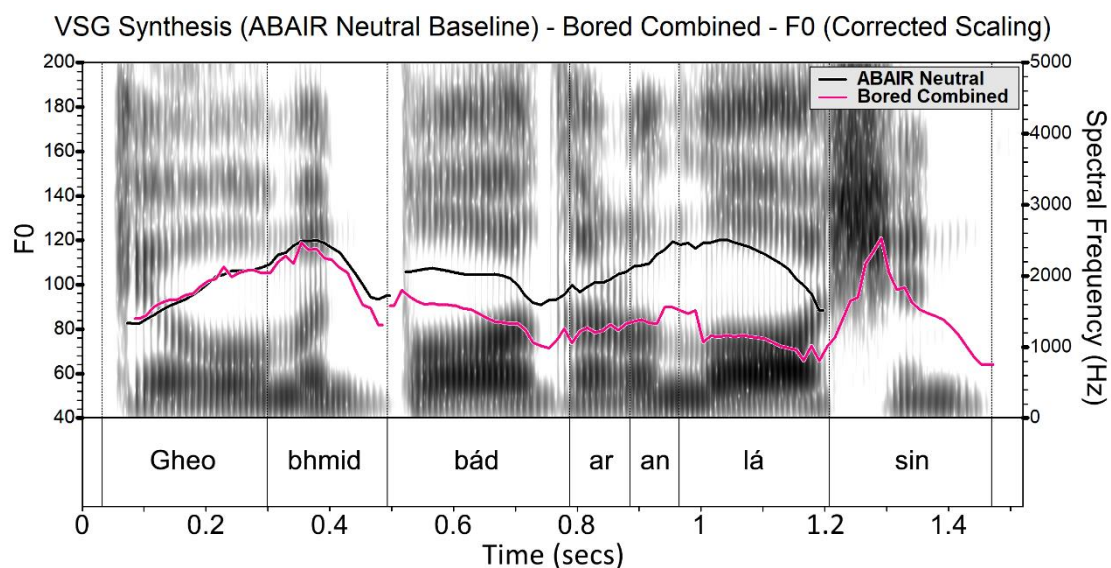


Figure D10 - F_0 contour of the combined modification of the rescaled ABAIR *neutral* stimulus to the *bored* target affect.

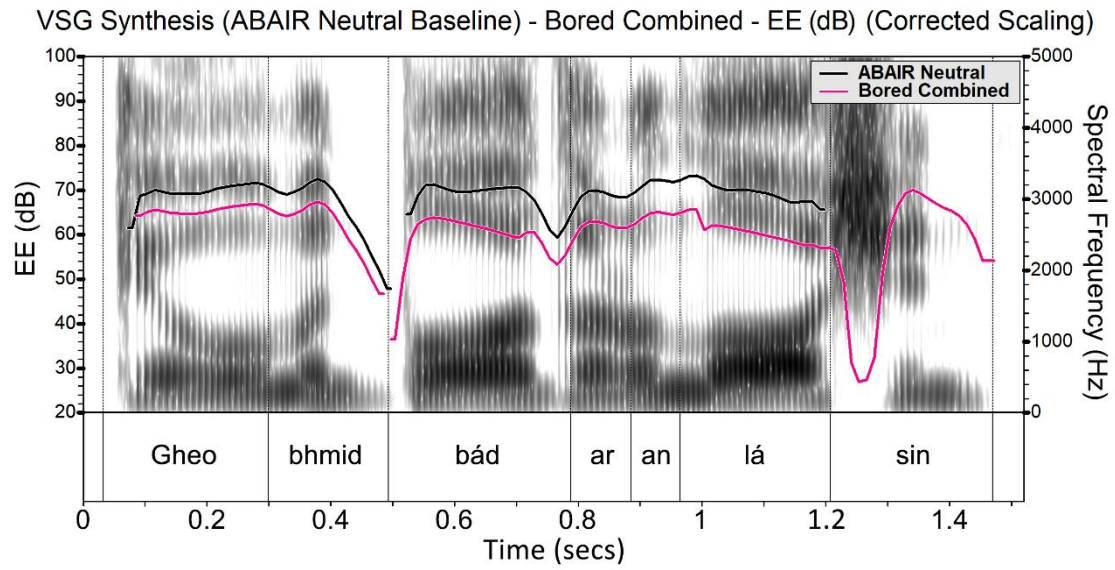


Figure D11 – EE contour of the combined modification of the rescaled ABAIR *neutral* stimulus to the *bored* target affect.

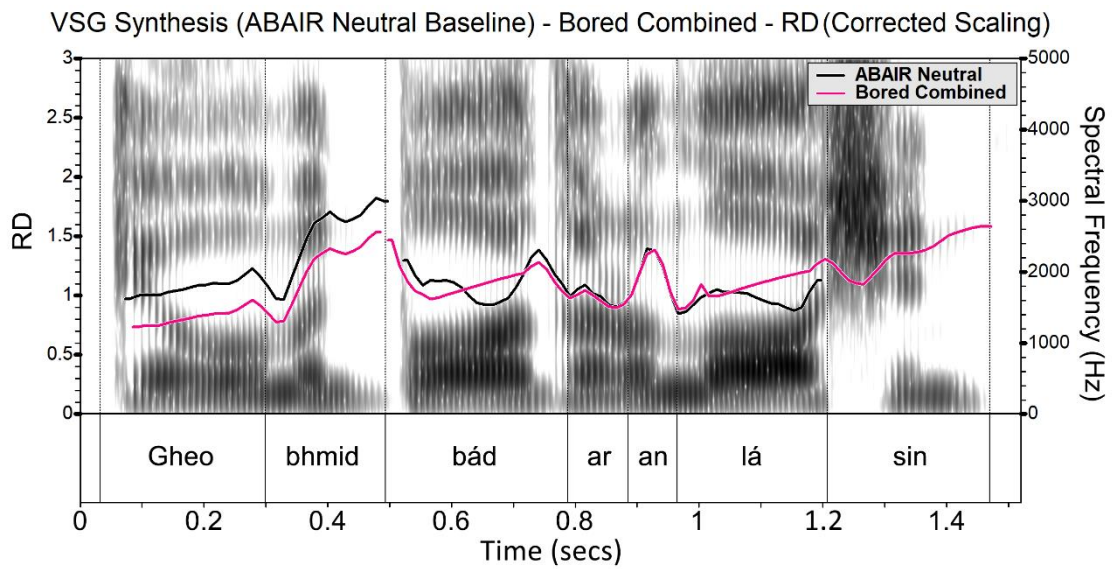


Figure D12 - RD contour of the combined modification of the rescaled ABAIR *neutral* stimulus to the *bored* target affect.

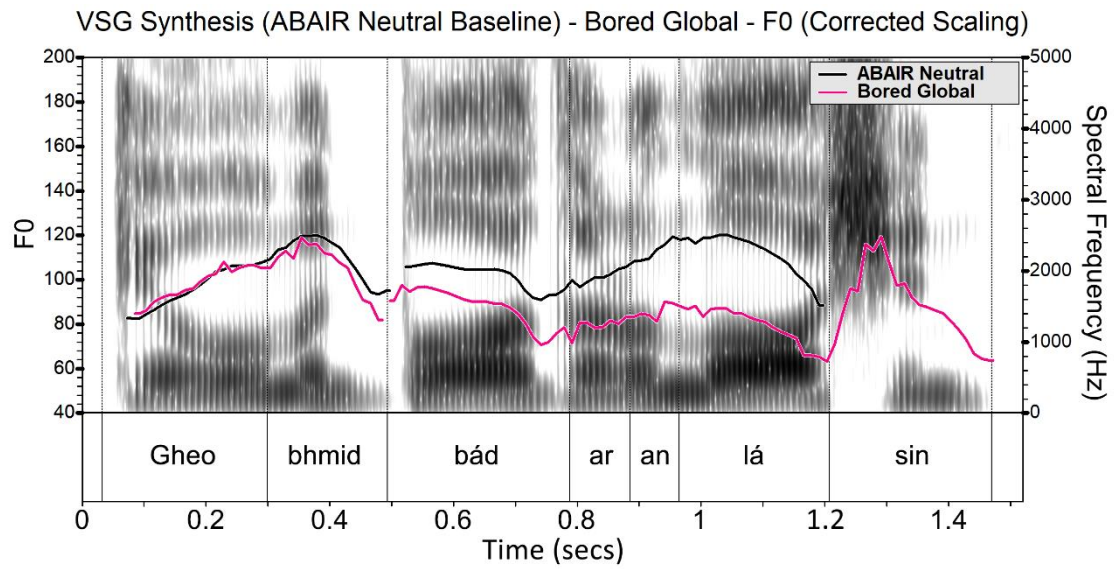


Figure D13 – F_0 contour of the global modification of the rescaled ABAIR *neutral* stimulus to the *bored* target affect.

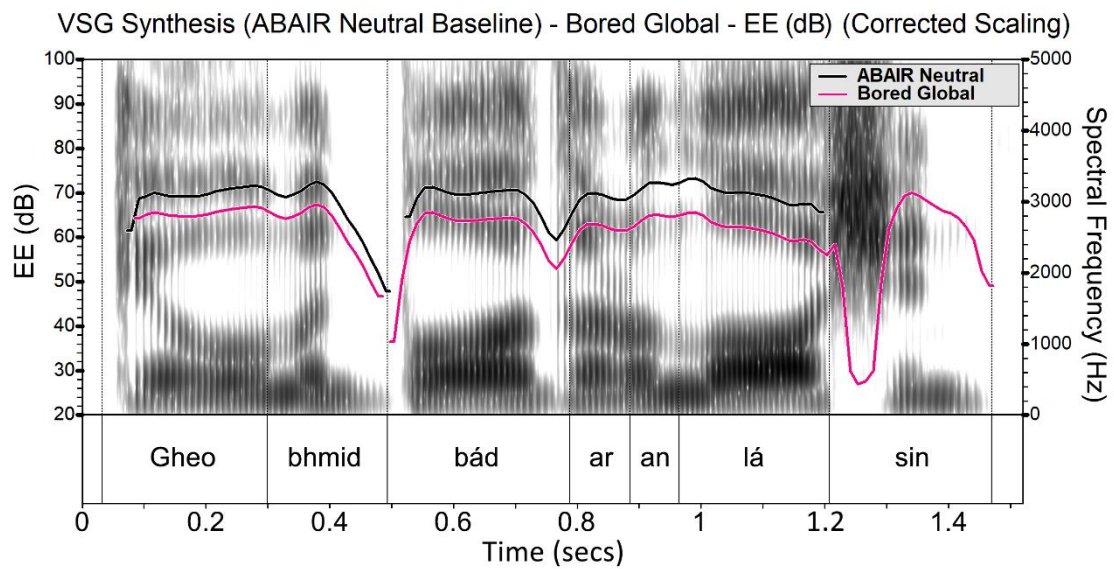


Figure D14 – EE contour of the global modification of the rescaled ABAIR *neutral* stimulus to the *bored* target affect.

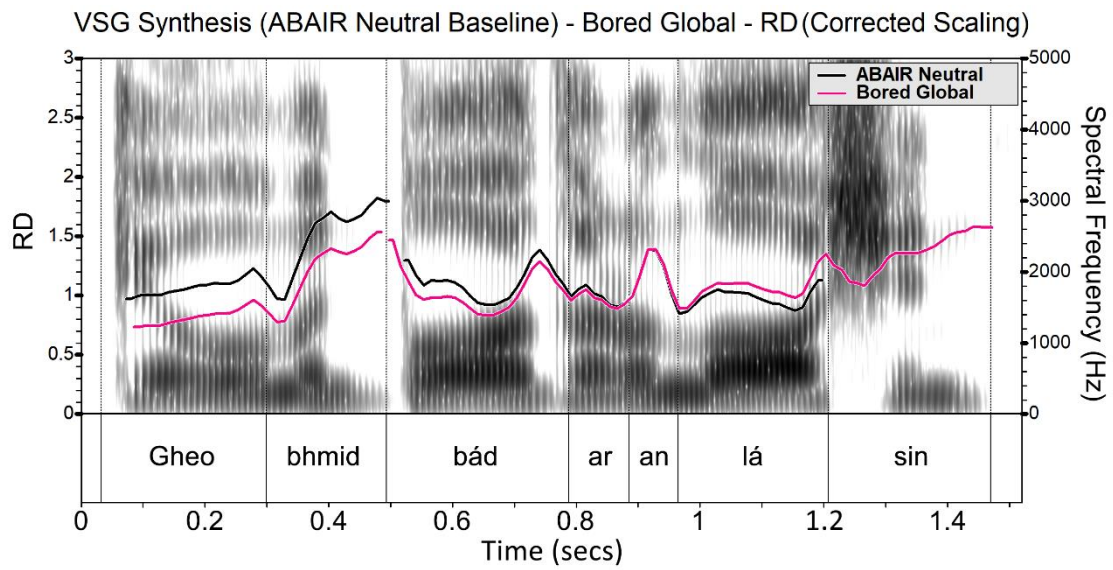


Figure D15 –RD contour of the global modification of the rescaled ABAIR *neutral* stimulus to the *bored* target affect.

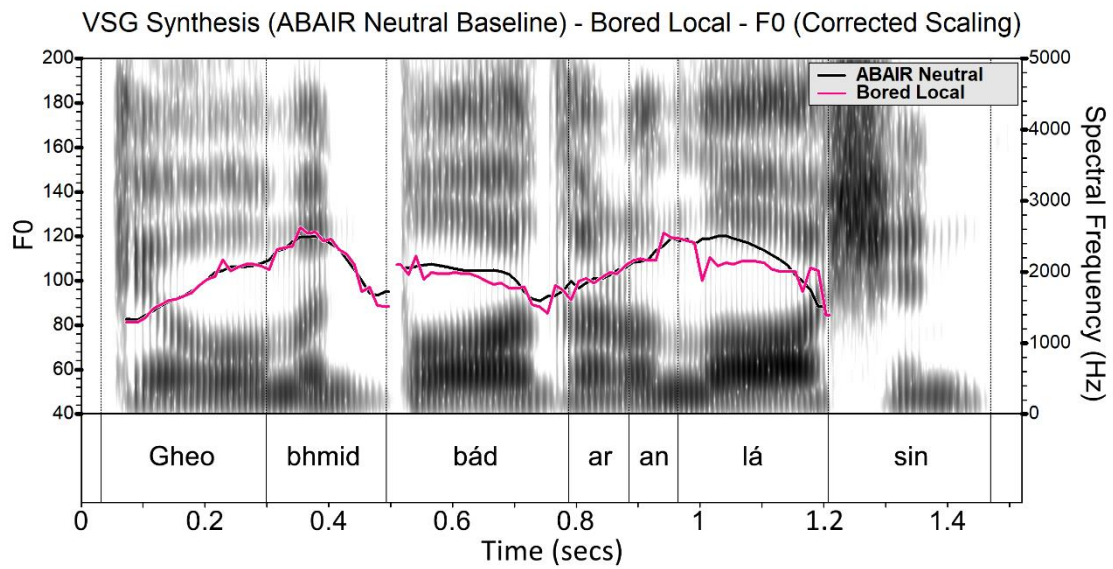


Figure D16 - F₀ contour of the local modification of the rescaled ABAIR *neutral* stimulus to the *bored* target affect.

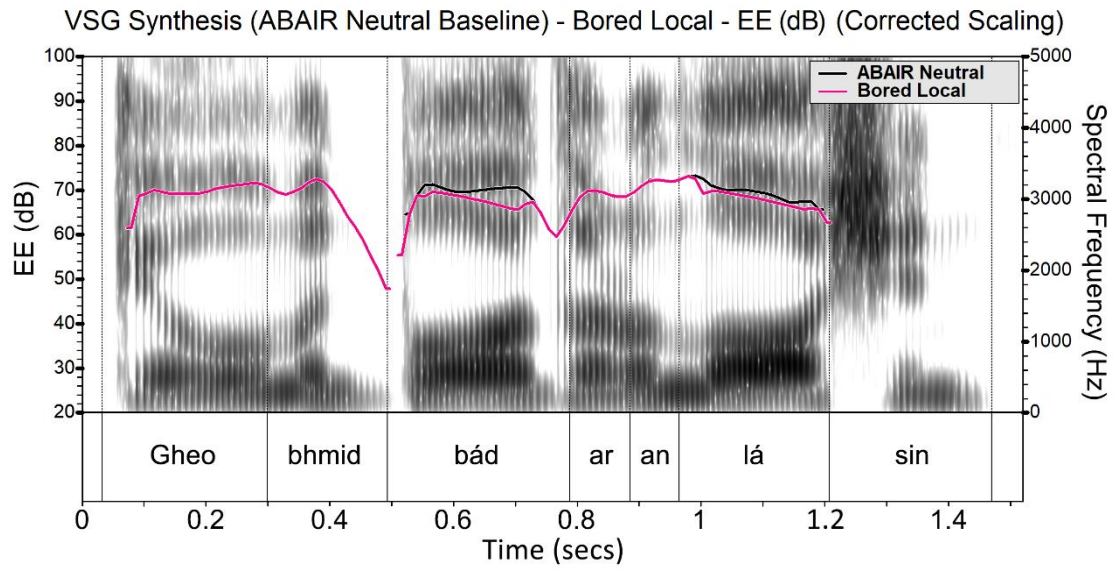


Figure D17 – EE contour of the local modification of the rescaled ABAIR *neutral* stimulus to the *bored* target affect.

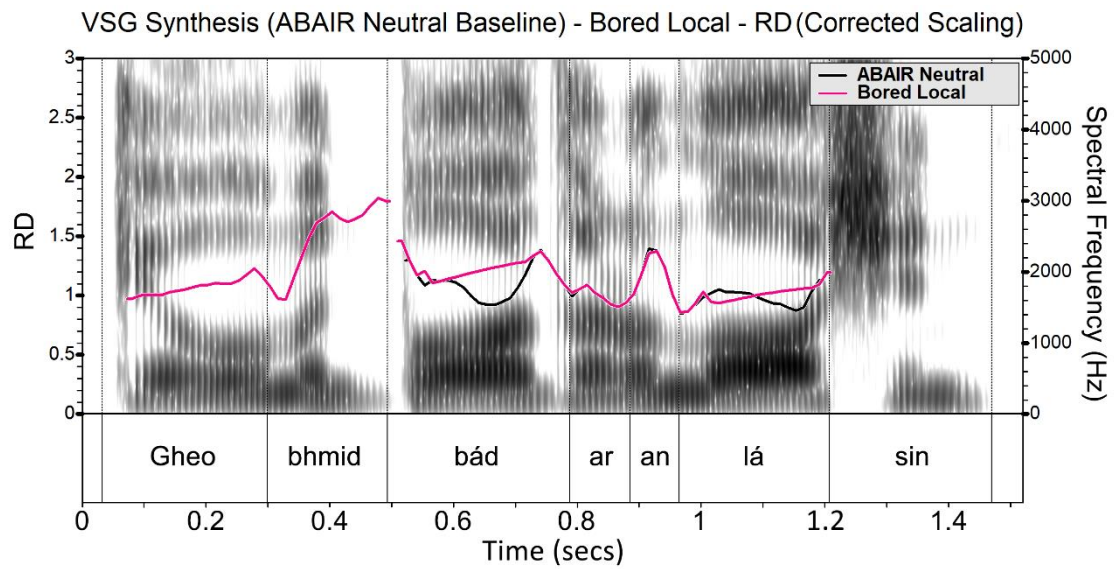


Figure D18 - RD contour of the local modification of the rescaled ABAIR *neutral* stimulus to the *bored* target affect.

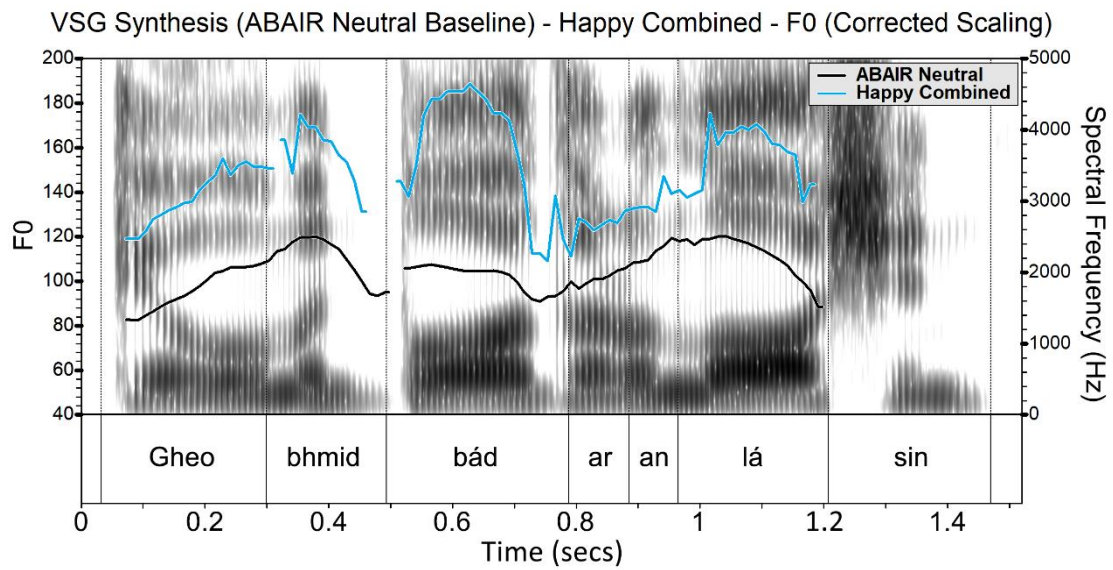


Figure D19 – F₀ contour of the combined modification of the rescaled ABAIR *neutral* stimulus to the *happy* target affect.

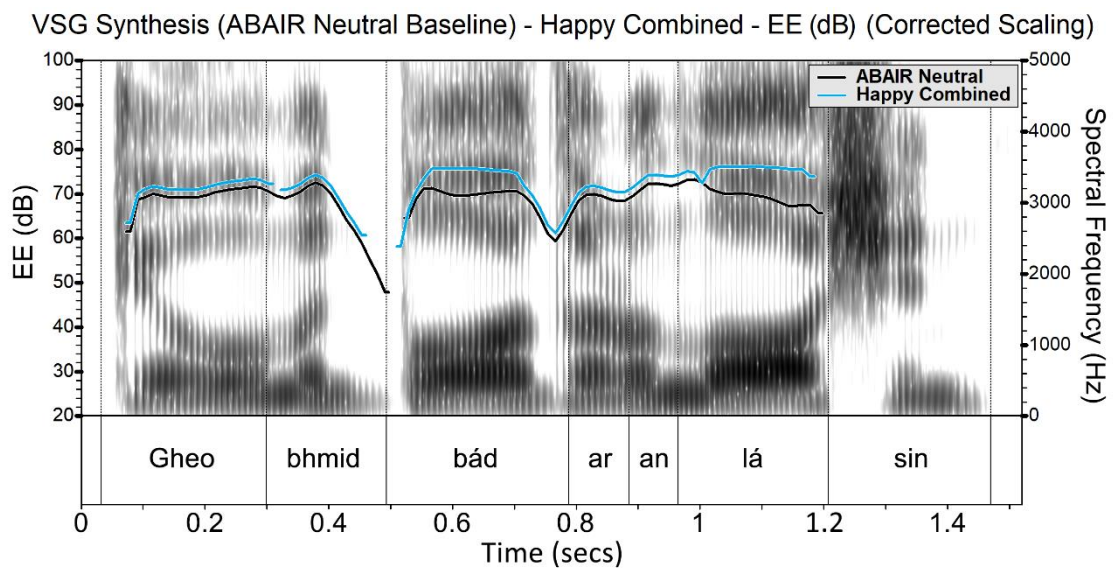


Figure D20 - EE contour of the combined modification of the rescaled ABAIR *neutral* stimulus to the *happy* target affect.

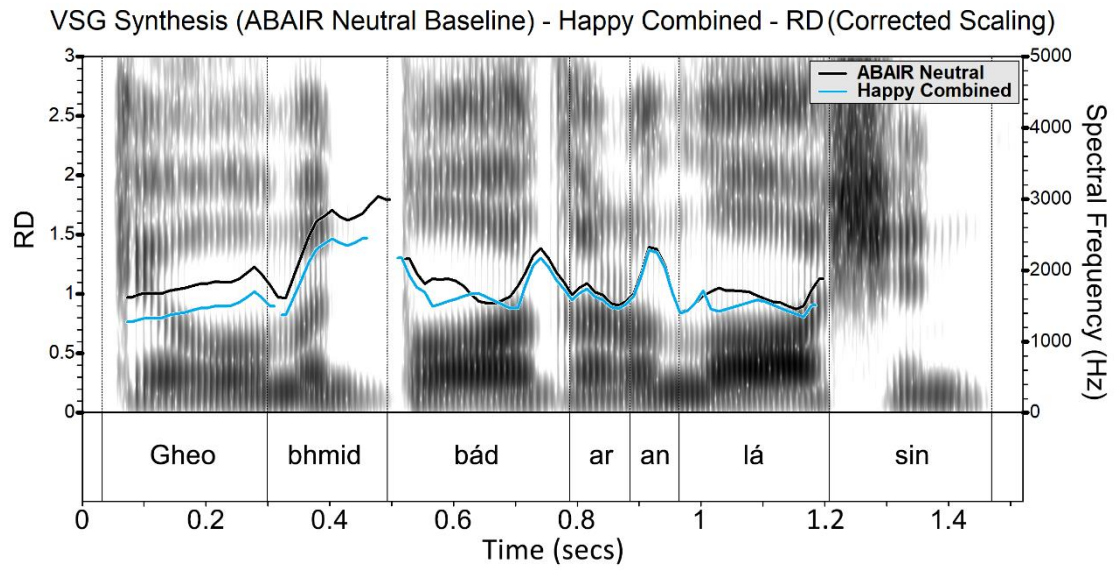


Figure D21 - RD contour of the combined modification of the rescaled ABAIR *neutral* stimulus to the *happy* target affect.

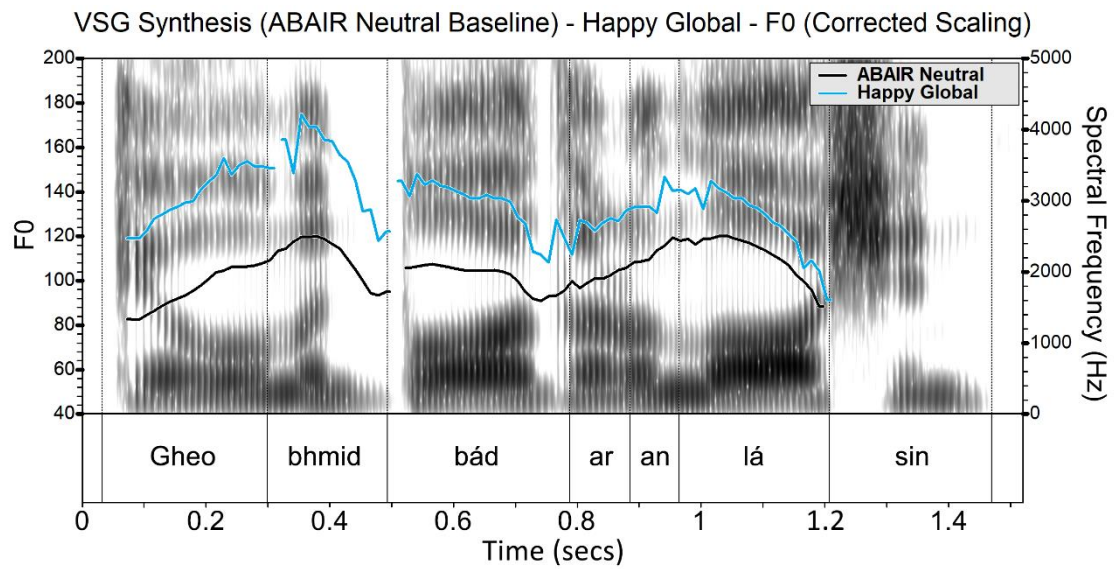


Figure D22 - F_0 contour of the global modification of the rescaled ABAIR *neutral* stimulus to the *happy* target affect.

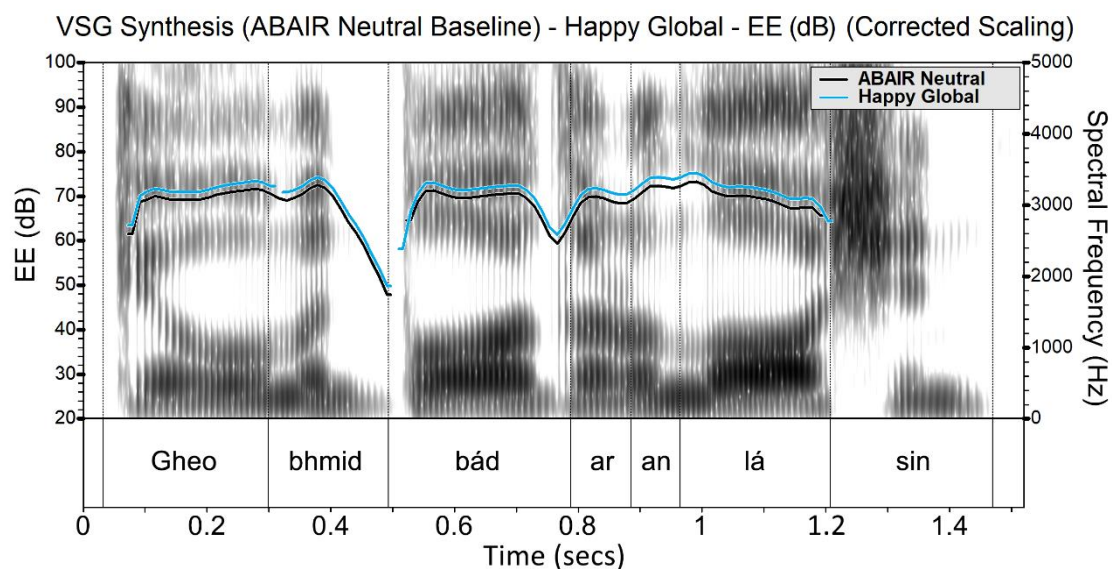


Figure D23 - EE contour of the global modification of the rescaled ABAIR *neutral* stimulus to the *happy* target affect.

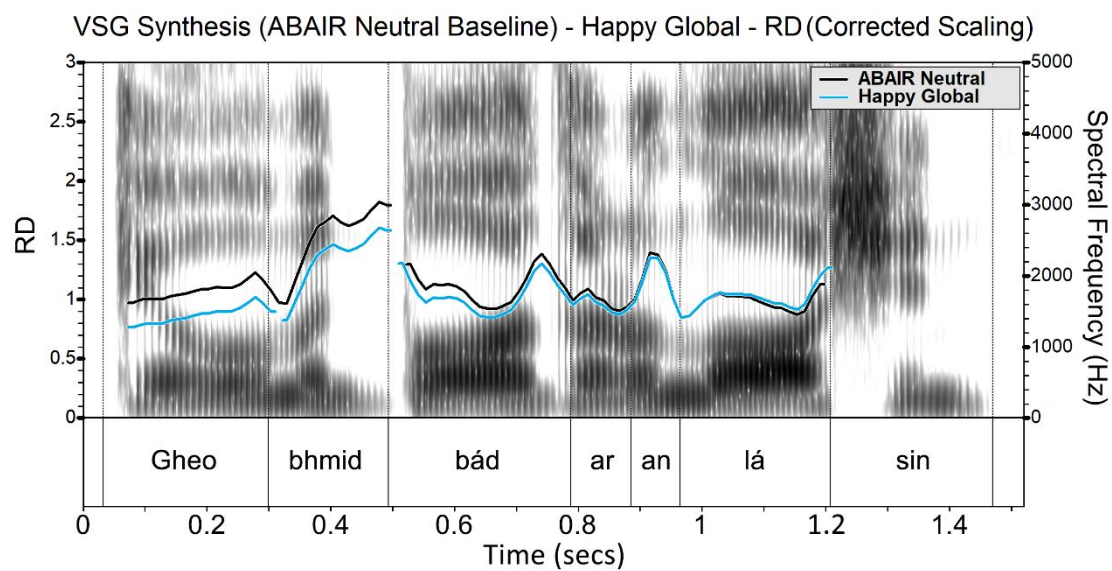


Figure D24 - RD contour of the global modification of the rescaled ABAIR *neutral* stimulus to the *happy* target affect.

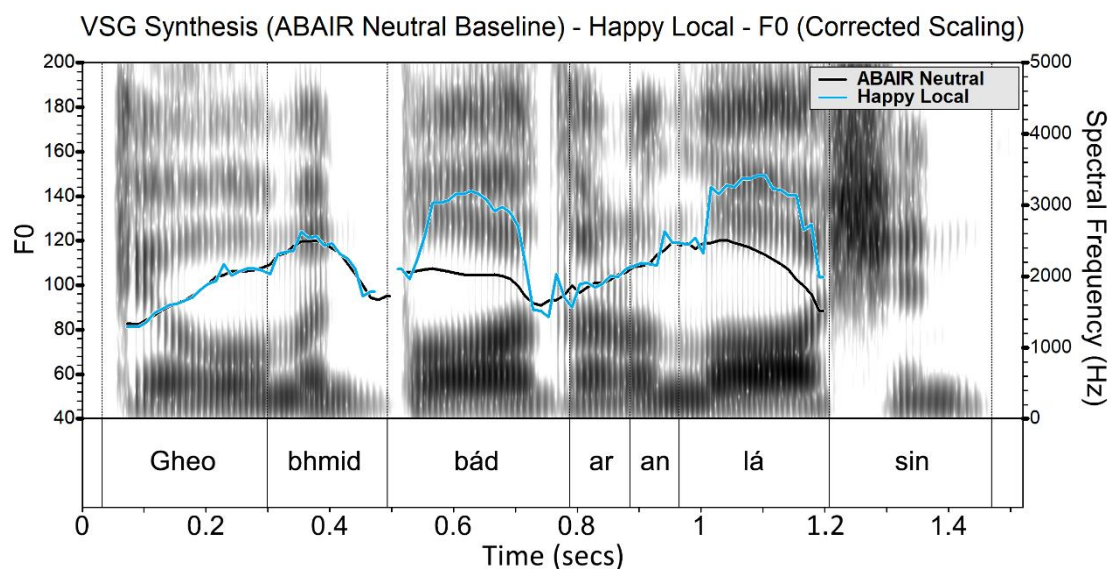


Figure D25 - F₀ contour of the local modification of the rescaled ABAIR *neutral* stimulus to the *happy* target affect.

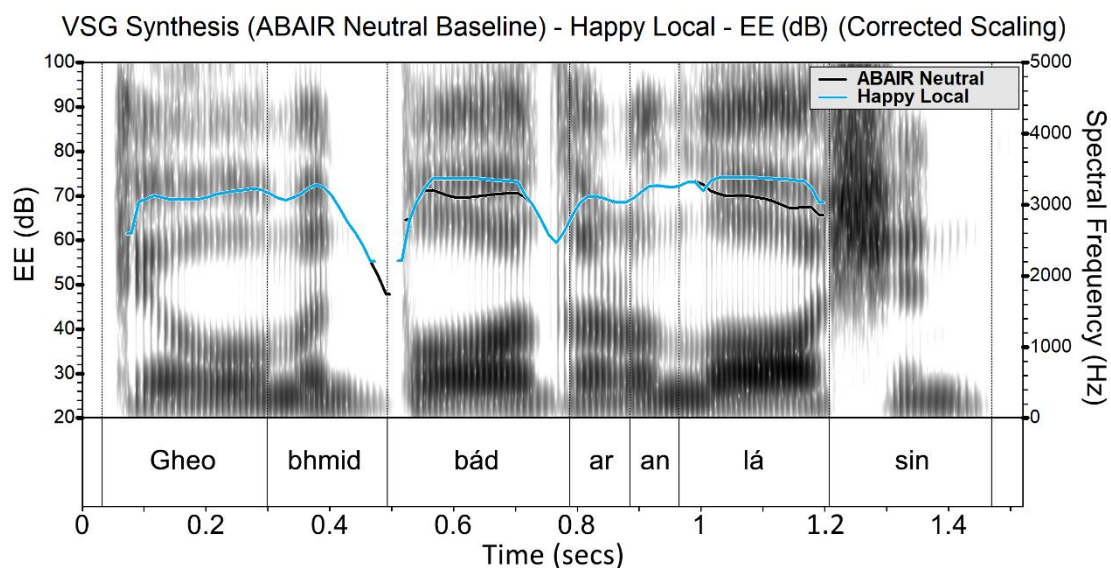


Figure D26 - EE contour of the local modification of the rescaled ABAIR *neutral* stimulus to the *happy* target affect.

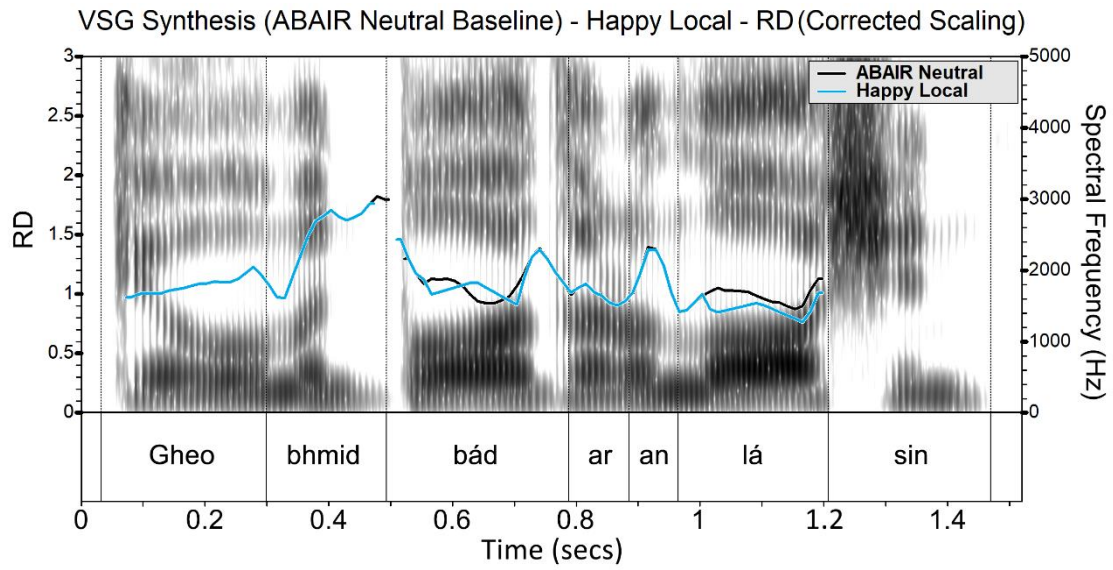


Figure D27 - RD contour of the local modification of the rescaled ABAIR *neutral* stimulus to the *happy* target affect.

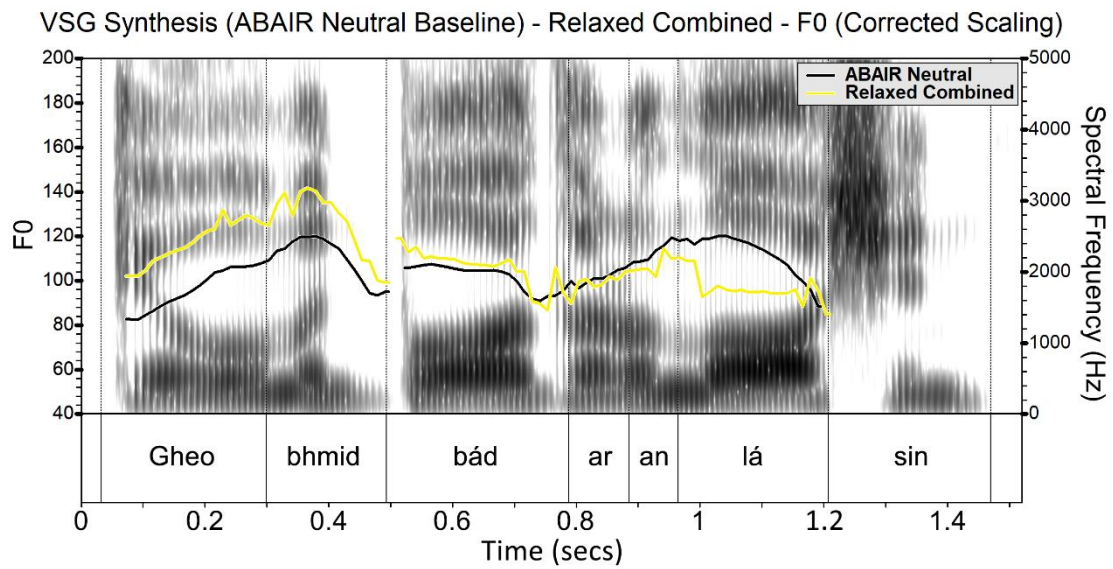


Figure D28 - F₀ contour of the combined modification of the rescaled ABAIR *neutral* stimulus to the *relaxed* target affect.

VSG Synthesis (ABAIR Neutral Baseline) - Relaxed Combined - EE (dB) (Corrected Scaling)

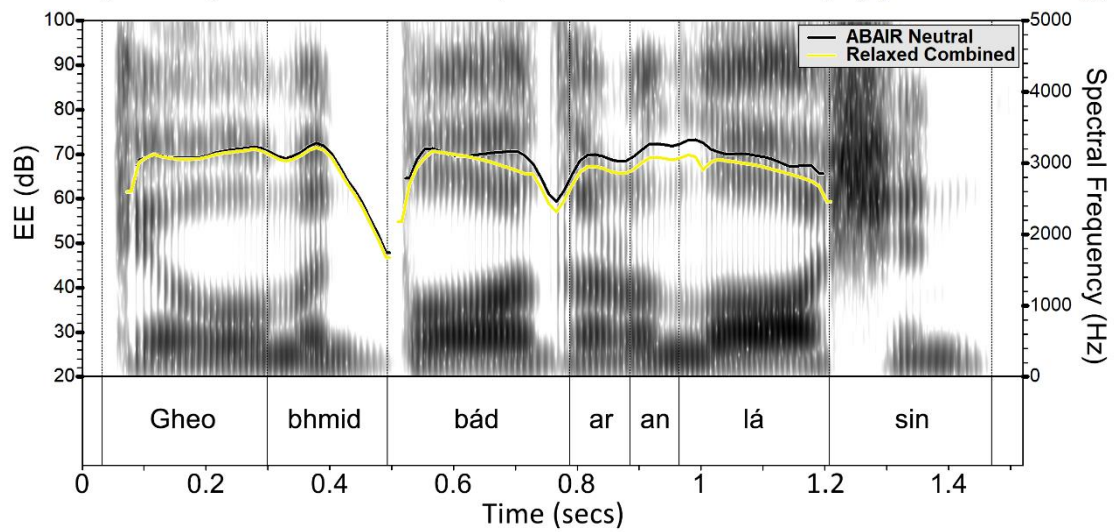


Figure D29 - EE contour of the combined modification of the rescaled ABAIR *neutral* stimulus to the *relaxed* target affect.

VSG Synthesis (ABAIR Neutral Baseline) - Relaxed Combined - RD(Corrected Scaling)

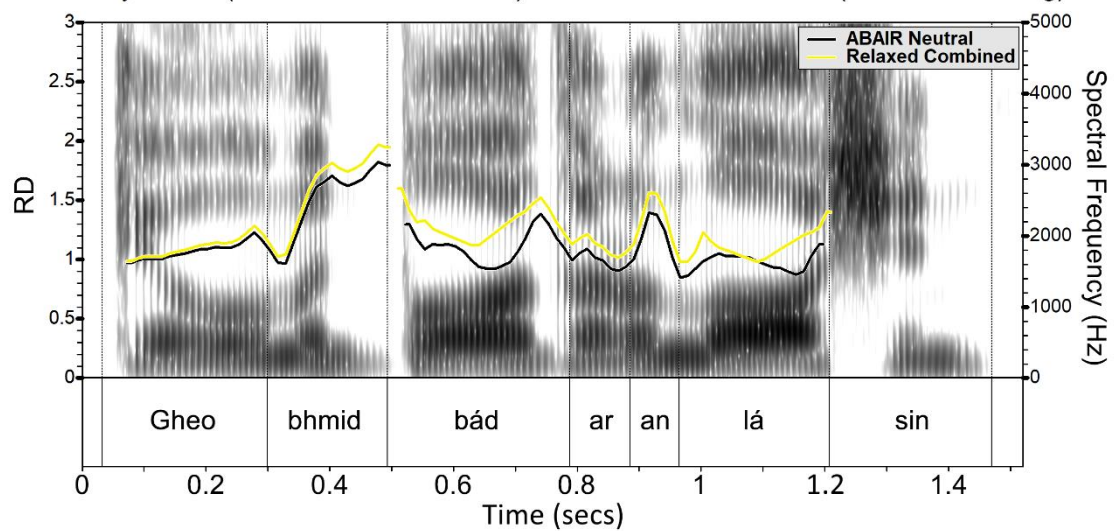


Figure D30 - RD contour of the combined modification of the rescaled ABAIR *neutral* stimulus to the *relaxed* target affect.

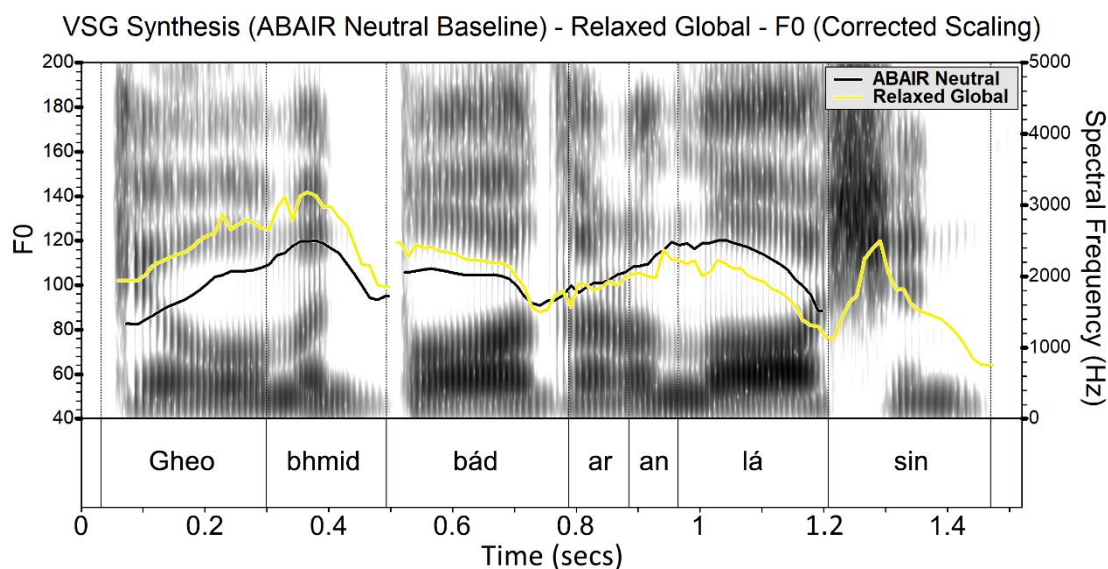


Figure D31 - F₀ contour of the global modification of the rescaled ABAIR *neutral* stimulus to the *relaxed* target affect.

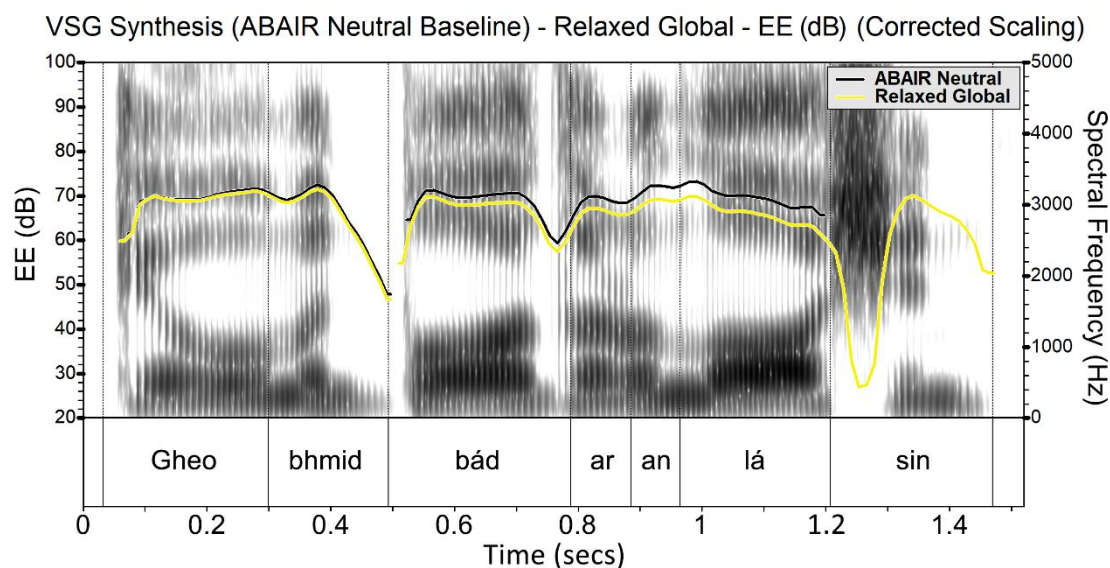


Figure D32 - EE contour of the global modification of the rescaled ABAIR *neutral* stimulus to the *relaxed* target affect.

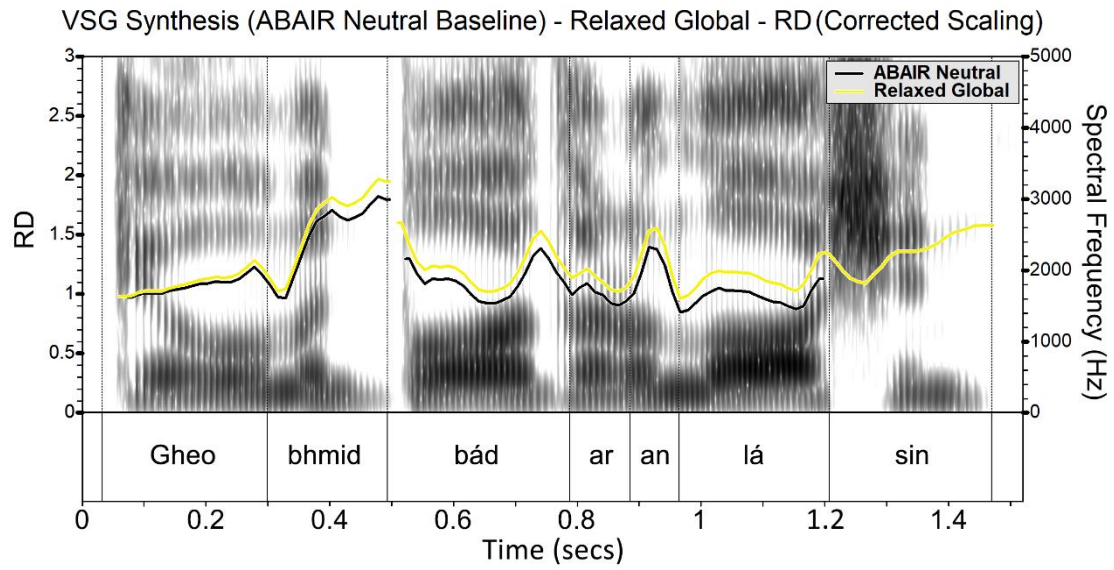


Figure D33 - RD contour of the global modification of the rescaled ABAIR *neutral* stimulus to the *relaxed* target affect.

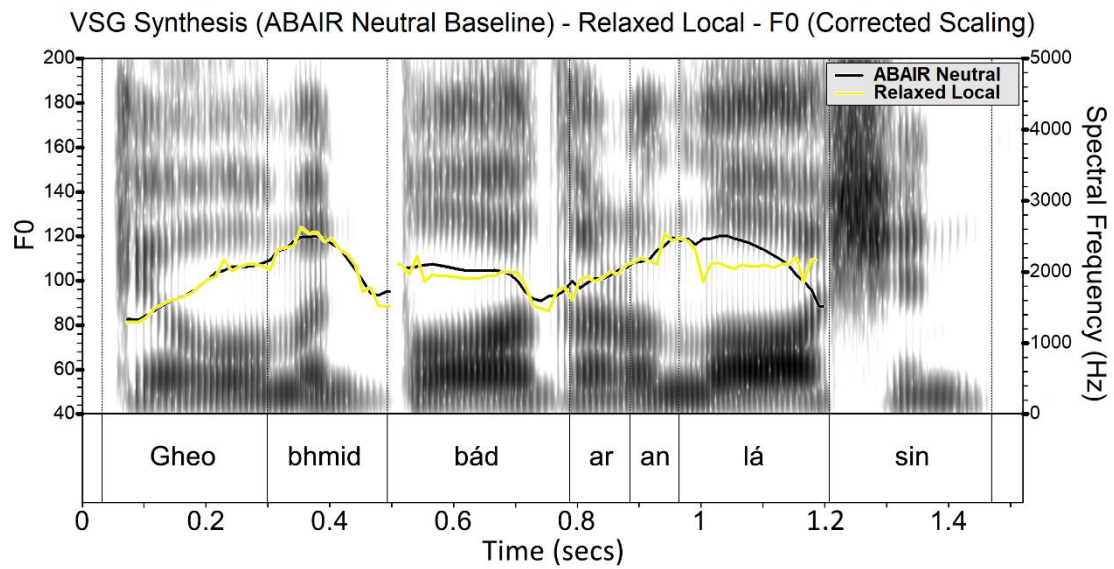


Figure D34 - F_0 contour of the local modification of the rescaled ABAIR *neutral* stimulus to the *relaxed* target affect.

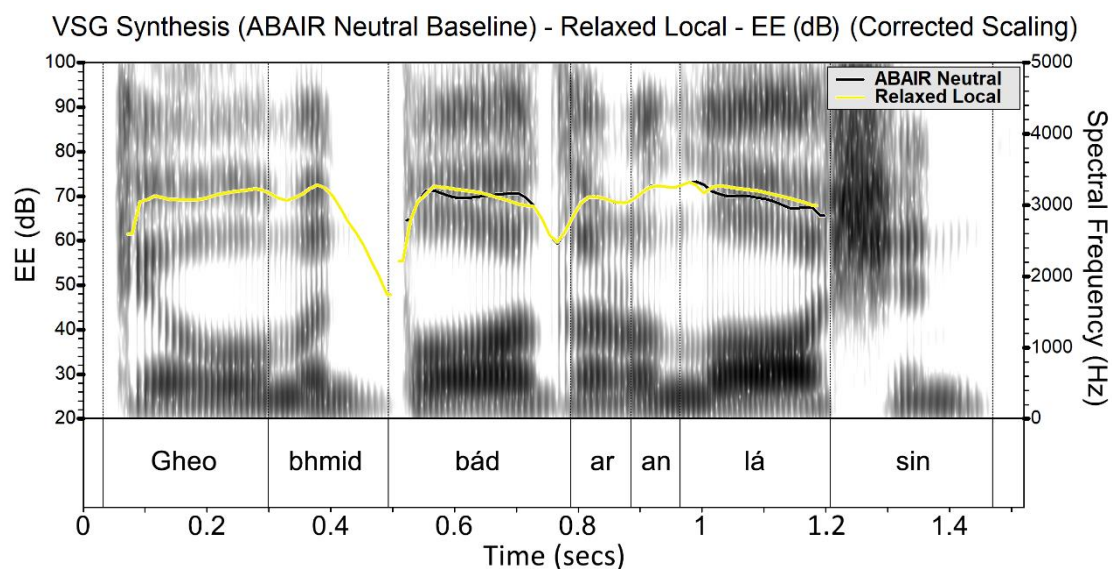


Figure D35 - EE contour of the global modification of the rescaled ABAIR *neutral* stimulus to the *relaxed* target affect.

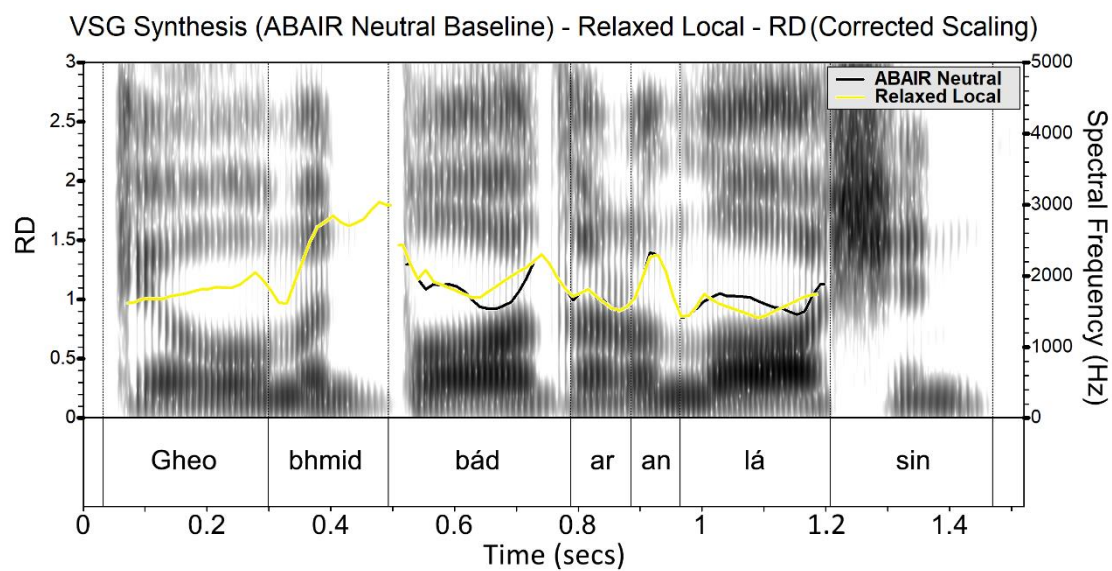


Figure D36 - RD contour of the local modification of the rescaled ABAIR *neutral* stimulus to the *relaxed* target affect.

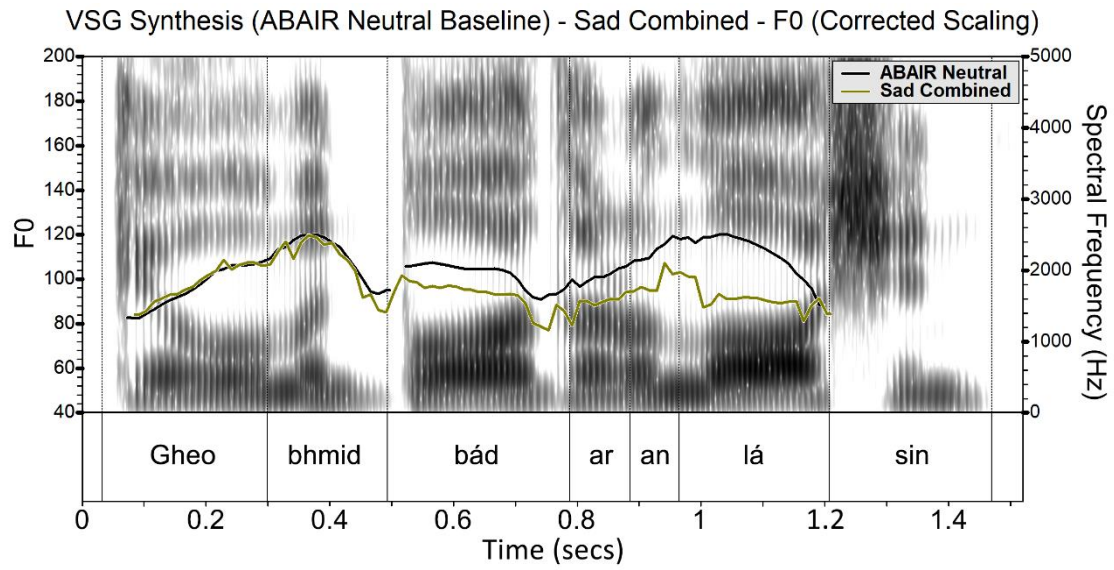


Figure D37 - F₀ contour of the combined modification of the rescaled ABAIR *neutral* stimulus to the *sad* target affect.

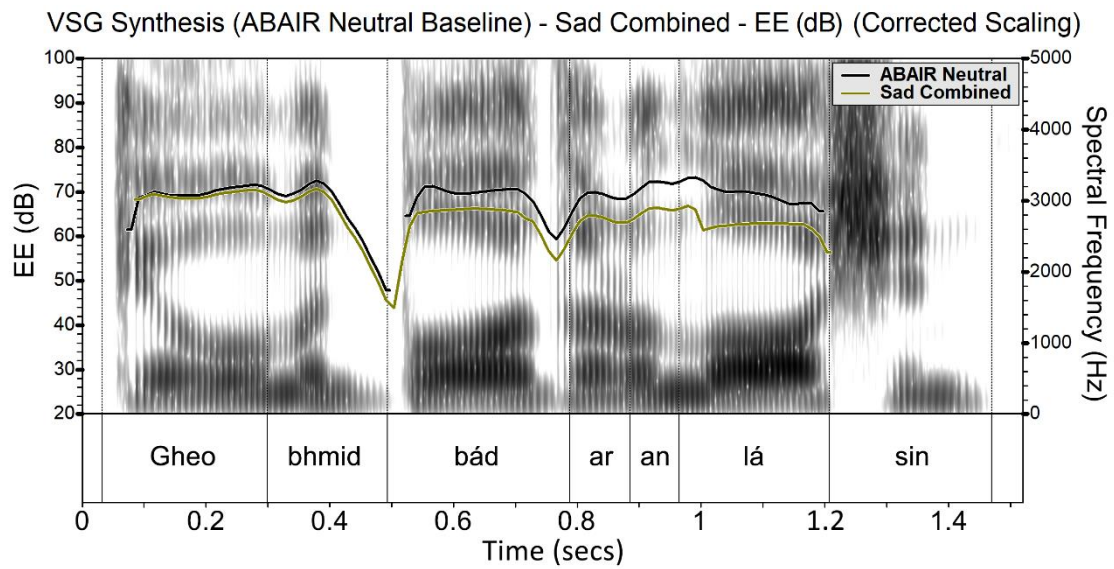


Figure D38 - EE contour of the combined modification of the rescaled ABAIR *neutral* stimulus to the *sad* target affect.

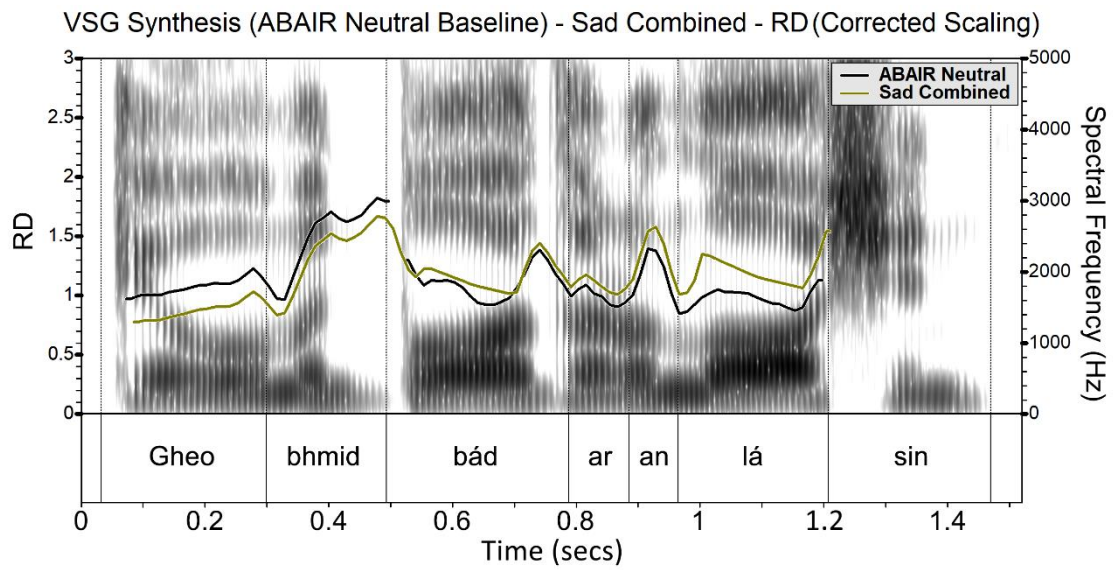


Figure D39 - RD contour of the combined modification of the rescaled ABAIR *neutral* stimulus to the *sad* target affect.

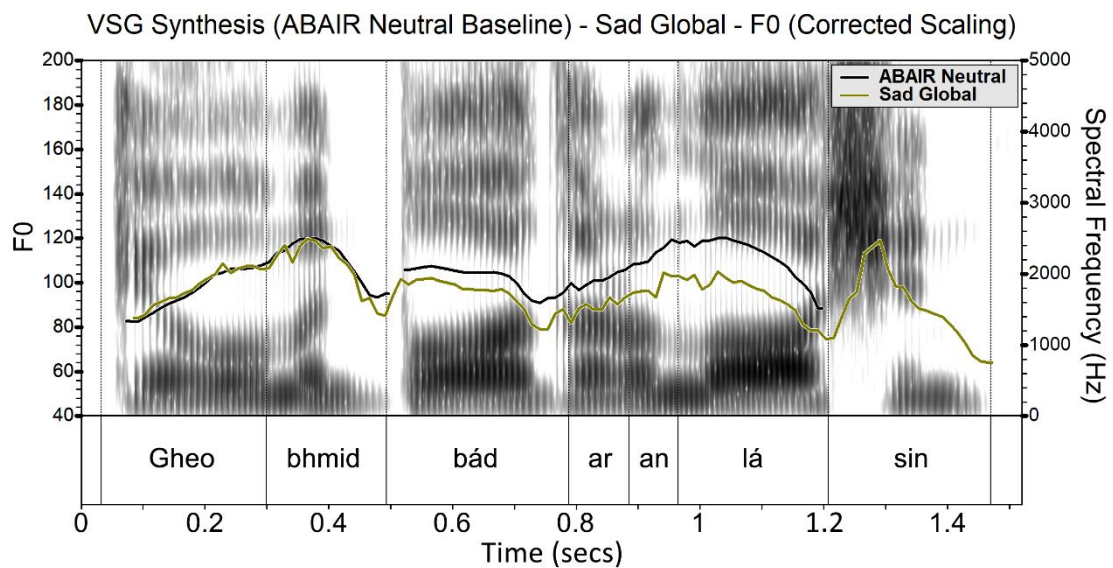


Figure D40 - F_0 contour of the global modification of the rescaled ABAIR *neutral* stimulus to the *sad* target affect.

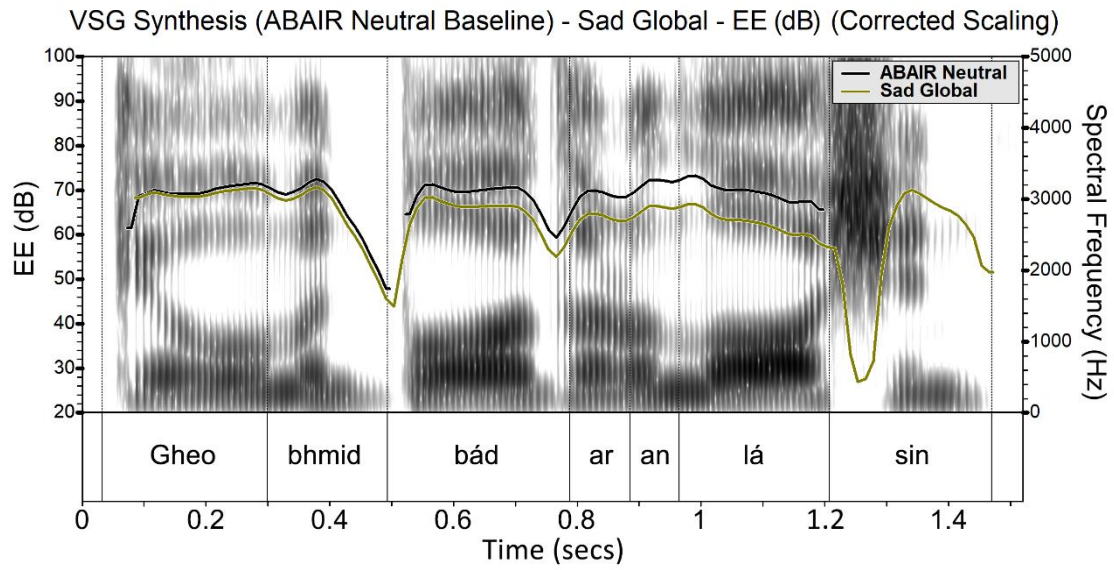


Figure D41 - EE contour of the global modification of the rescaled ABAIR *neutral* stimulus to the *sad* target affect.

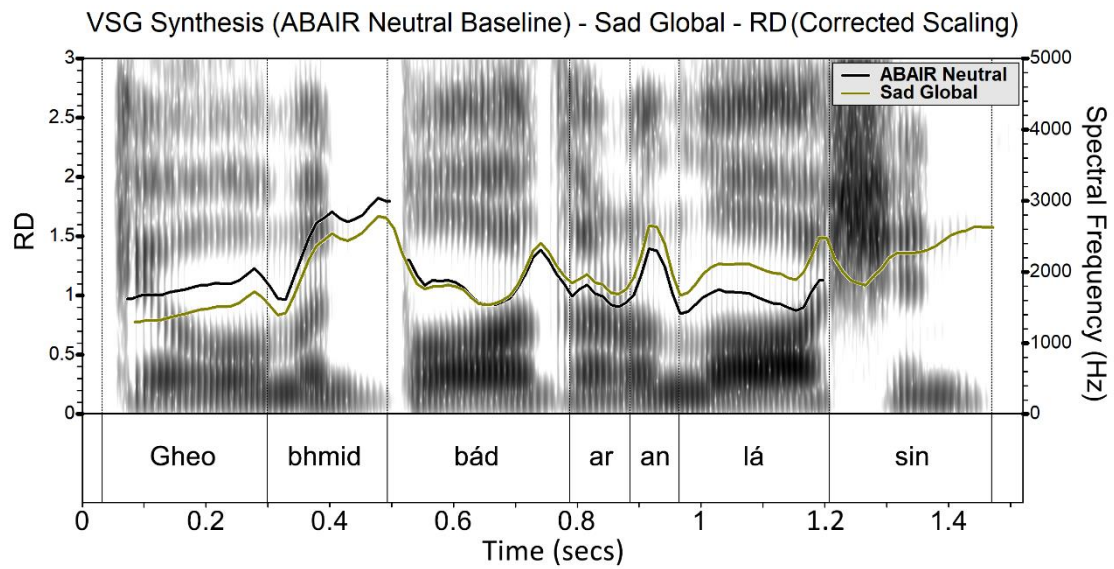


Figure D42 - RD contour of the combined modification of the rescaled ABAIR *neutral* stimulus to the *sad* target affect.

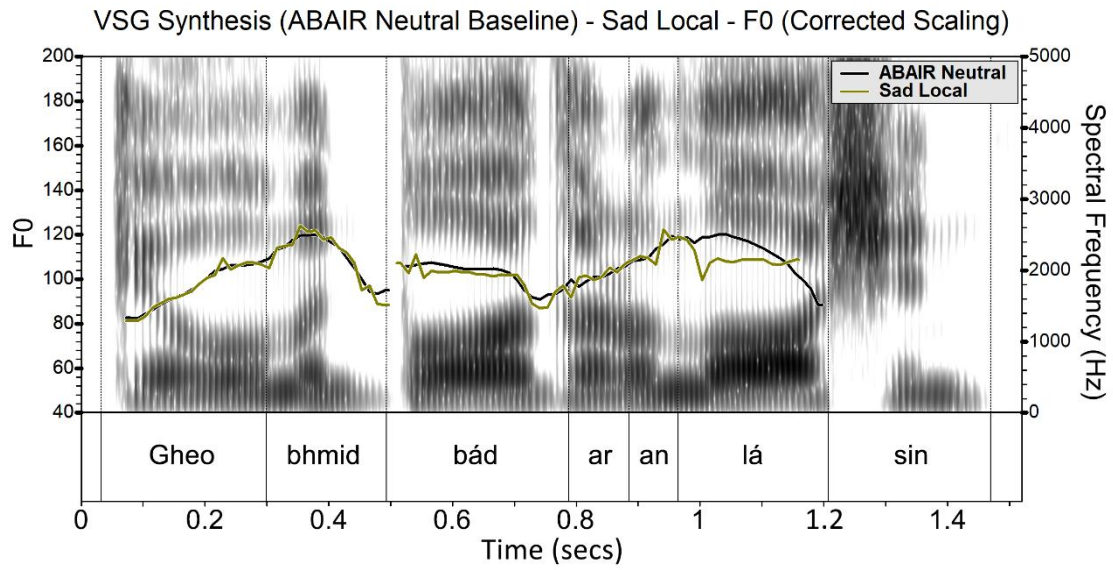


Figure D43 - F₀ contour of the local modification of the rescaled ABAIR *neutral* stimulus to the *sad* target affect.

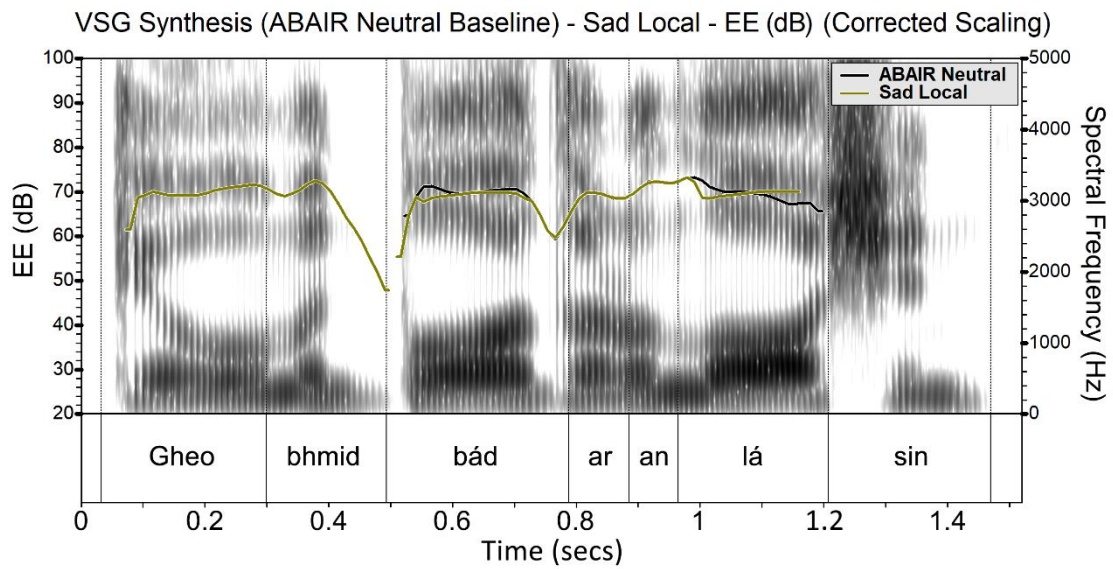


Figure D44 – EE contour of the local modification of the rescaled ABAIR *neutral* stimulus to the *sad* target affect.

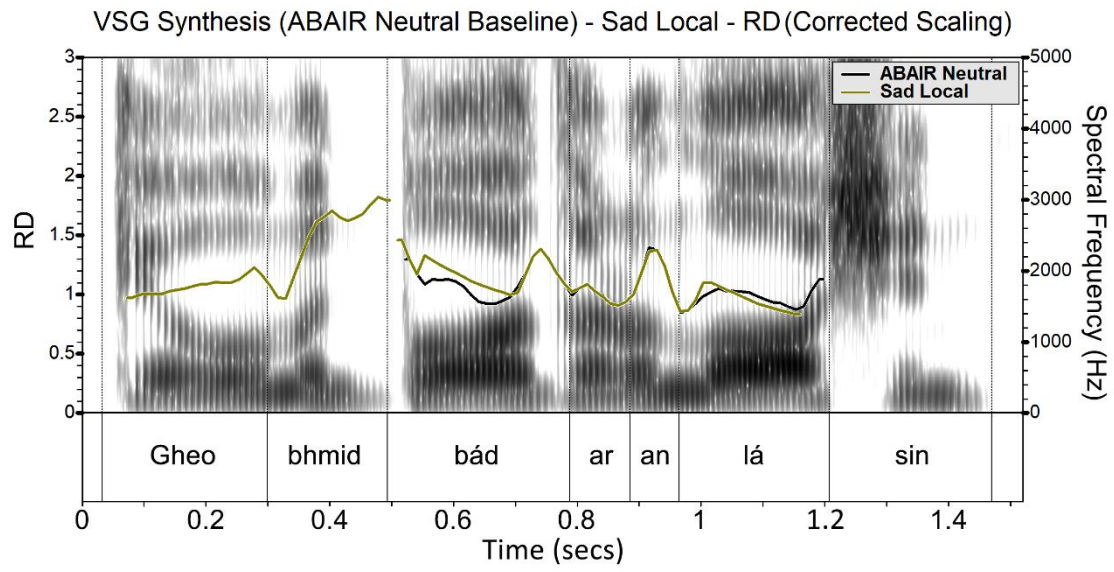


Figure D45 - RD contour of the local modification of the rescaled ABAIR *neutral* stimulus to the *sad* target affect.

Appendix E

Still frames of all facial expressions and intensities for the visual context experiments.

Original perception tests with videos and synthesized stimuli combinations can be found in the links below:

Experiment 03 (Cross-activation Incongruence):

https://docs.google.com/forms/d/e/1FAIpQLScAiJeGhgg_r7eLfoNgyj48F2XLdWgyA-qplriNWB97tJxw/viewform?usp=sf_link

Experiment 04 (Same Activation Incongruence):

https://docs.google.com/forms/d/e/1FAIpQLSdzxbXbsikmldpdLoUyBH2fXH24WuFV6hq_4axL3FjotdxqAQ/viewform?usp=sf_link



Figure E1 – Neutral Visual Stimulus.



Figure E2 – Angry (Intense) Visual Stimulus.



Figure E3 – Angry (Moderate) Visual Stimulus.



Figure E4 – Angry (Slight) Visual Stimulus.



Figure E5 – Bored (Intense) Visual Stimulus.



Figure E6 – Bored (Moderate) Visual Stimulus.



Figure E7 – Bored (Slight) Visual Stimulus.

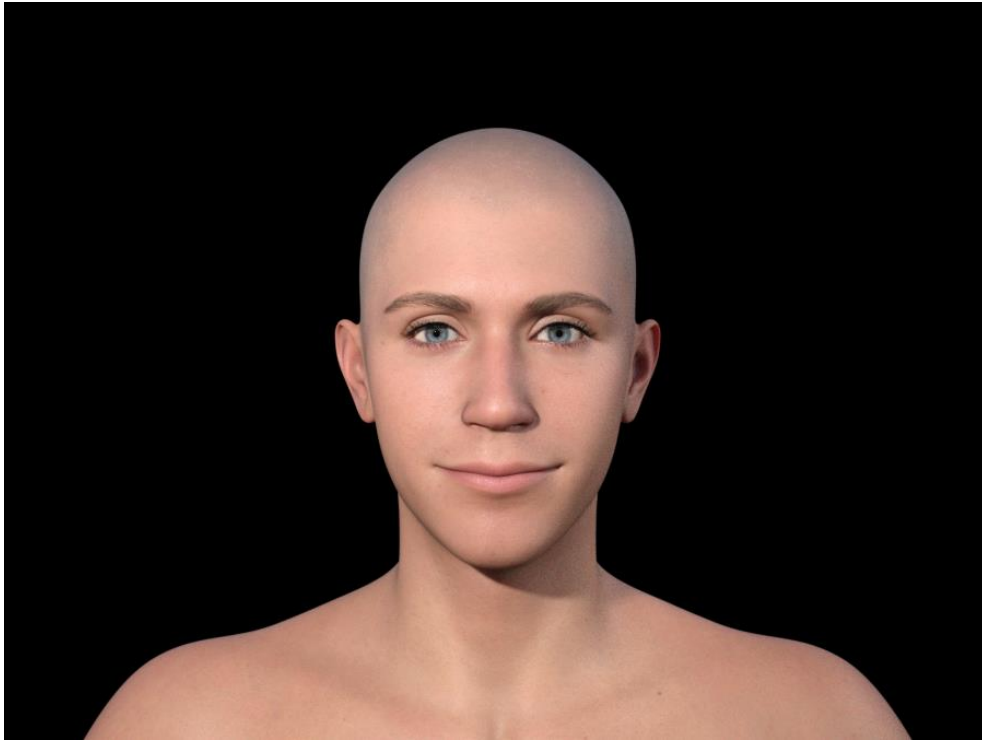


Figure E8 – Happy (Intense) Visual Stimulus.

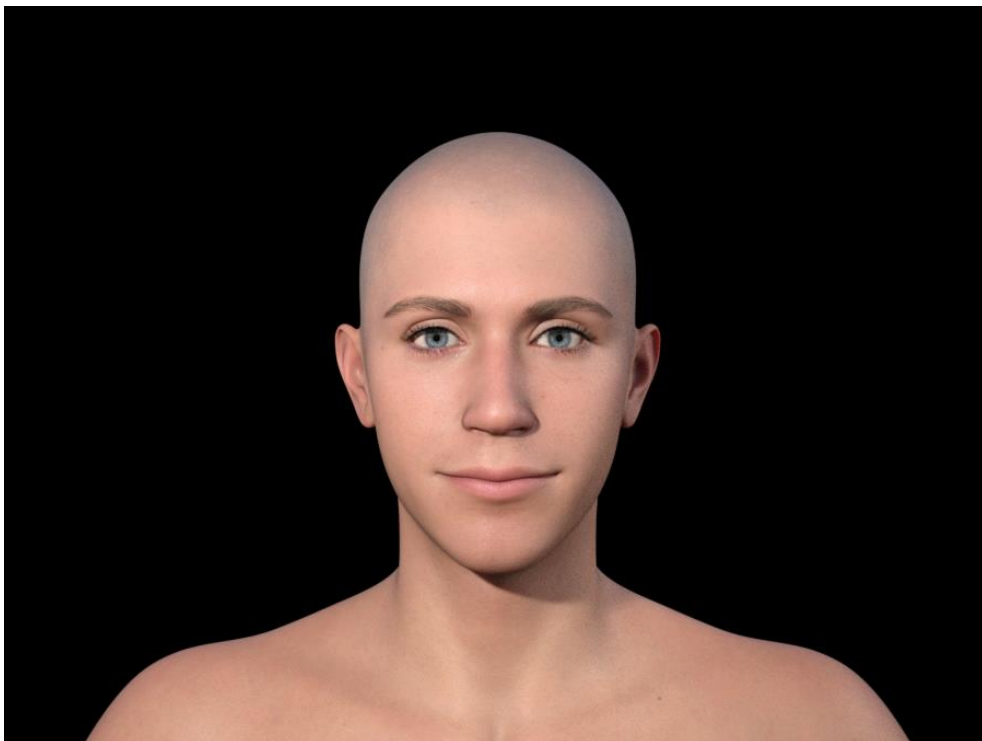


Figure E9 – Happy (Moderate) Visual Stimulus.

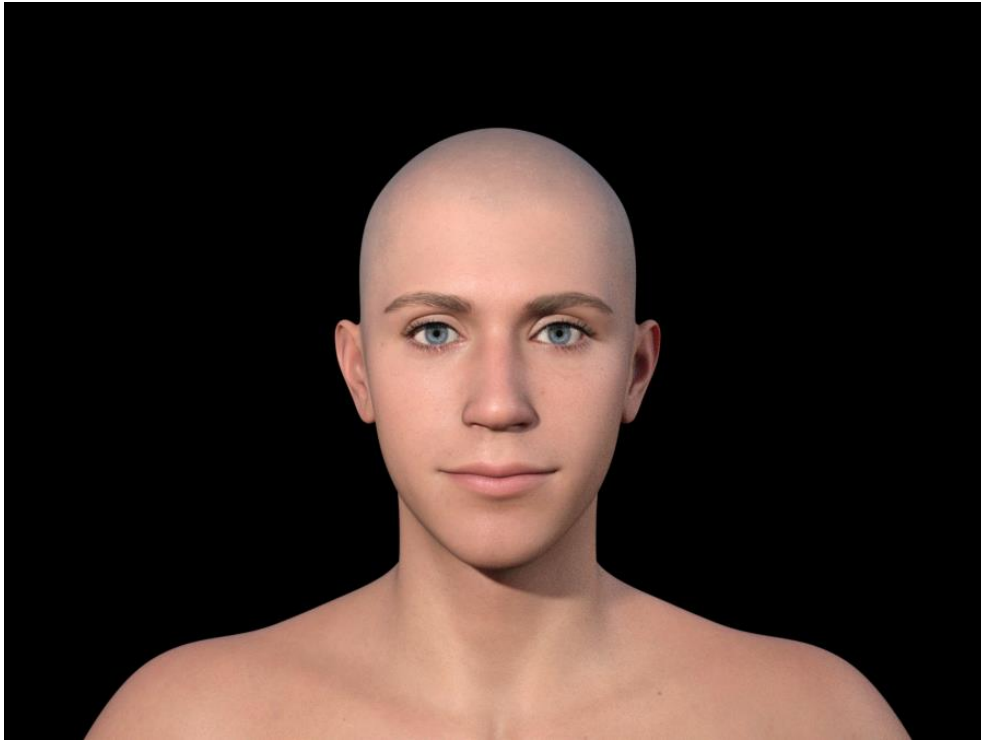


Figure E10 – Happy (Slight) Visual Stimulus.



Figure E11 – Interested (Intense) Visual Stimulus.

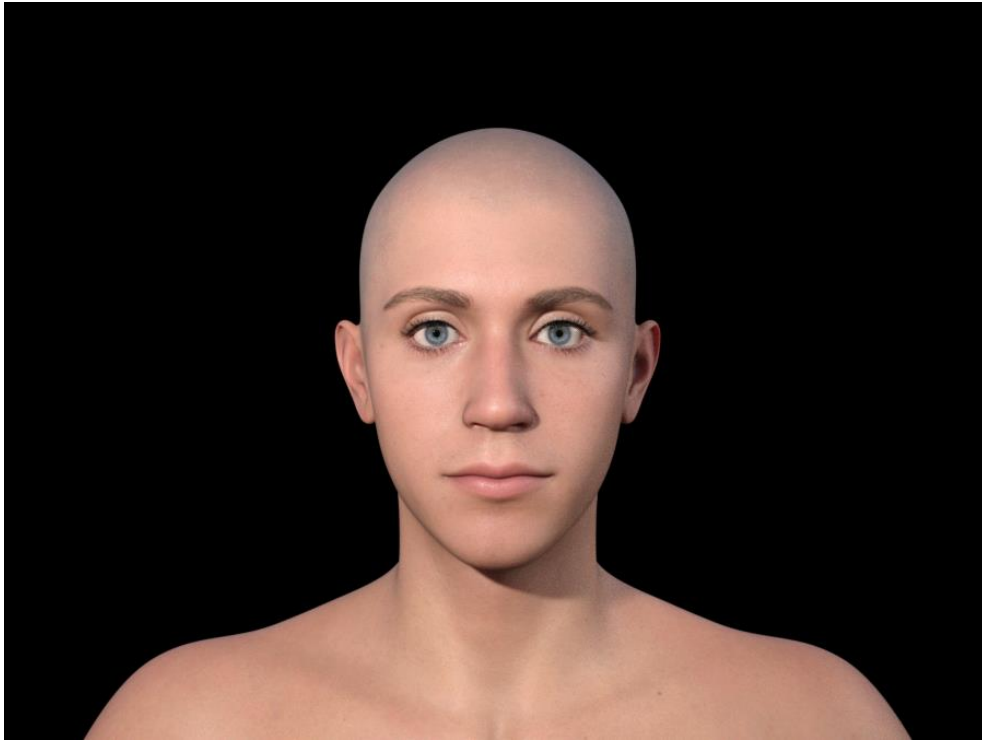


Figure E12 – Interested (Moderate) Visual Stimulus.



Figure E13 – Interested (Slight) Visual Stimulus.

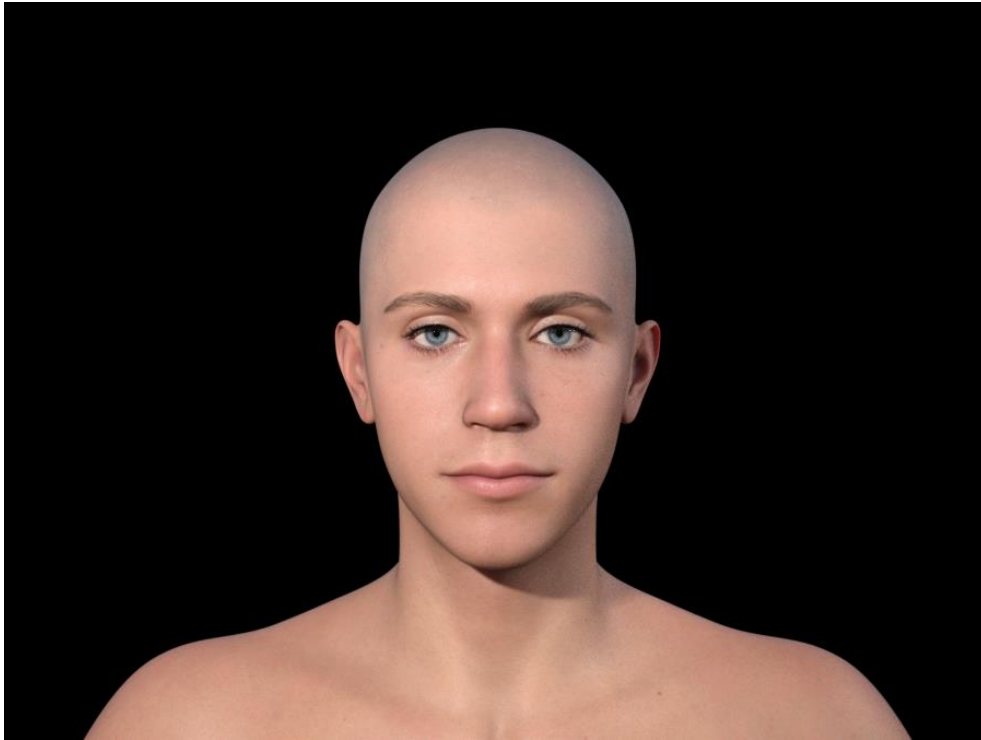


Figure E14 – Relaxed (Intense) Visual Stimulus.



Figure E15 – Relaxed (Moderate) Visual Stimulus.



Figure E16 – Relaxed (Slight) Visual Stimulus.



Figure E17 – Sad (Intense) Visual Stimulus.



Figure E18 – Sad (Moderate) Visual Stimulus.



Figure E19 – Sad (Slight) Visual Stimulus.