Exploring the prosody of affective speech

Anna M. Giovannini, Ailbhe Ní Chasaide, Christer Gobl

School of Linguistic, Speech and Communication Sciences, Trinity College Dublin, Ireland

https://doi.org/10.36505/ExLing-2022/13/0019/000561

Abstract

This paper introduces a research project on voice quality and affect expression. It explores affective prosody by investigating the relationship between voice source parameter changes and perceived affect. Firstly, it aims to examine the relative contribution of voice source shifts occurring globally across an utterance and shifts that are aligned to the prosodic structure of the utterance. Secondly, it aims to formulate a simple model for affect expression that could, in principle, be applied to text-to-speech synthesis systems for Irish (Gaelic) dialects. The analytic methods to be used include voice source and intonation analysis of utterances produced to portray a range of emotions, and perception experiments with stimuli varying in terms of global vs. local, structured source manipulations.

Keywords: emotion, synthesis, prosody, voice source, affect

Introduction

Current linguistic prosodic research largely avoids analysis of affective prosody. In fact, most research in the area of affective speech has been conducted by psychologists such as Scherer and colleagues (Scherer, 2003; Juslin and Scherer, 2005). However, although there is an emerging consensus that voice quality is central to the communication of affect, most analyses have tended to focus on global shifts in fundamental frequency (f0), intensity and tempo, while the crucial parameters of voice quality (tone of voice) are largely absent (but see, for example, Gobl and Ní Chasaide, 2003). Furthermore, beyond global average values, there is little account of whether and to what extent the changes to the voice source are in fact global shifts affecting entire utterances, or within-utterance shifts that take account of their known prosodic structure.

Building on earlier research on voice quality and affect expression (e.g., Murphy, Yanushevskaya, Ní Chasaide and Gobl, 2022), the present project sets out to explore affective prosody through analysis, and ultimately through synthesis, to investigate the relationship between voice source parameters (including f0) and perceived affect. A first objective is to examine to what extent affective changes may be cued by global, utterance-wide shifts in the source parameters and/or local shifts that are aligned to the prosodic structure of the utterance.

The second goal of this research is to build a simple model for affect expression that can, in principle, be exploited in speech synthesis. This goal is motivated by practical considerations, as it would be desirable to be able to implement basic affective shifts in the synthetic speech output of an Irish (Gaelic) text-to-speech (TTS) system (www.abair.ie) which is increasingly being used in applications for education and for users with disabilities. We aspire to enable a synthetic utterance to be produced with some basic affective modulation, e.g., to render the narration of a story more engaging to a young user, or to allow a disabled user to modify the voice of speech-based systems they use to communicate.

The research question for this project would therefore be:

Research questions

What are the roles of global (utterance-wide) and local (within-utterance, prosodically structured) changes in voice parameters (i.e., involving voice quality and fundamental frequency) in signalling affective states?

Hypothesis

The signaling of affect is ultimately more a matter of prosodically structured shifts in the voice source parameters than global shifts through the entirety of an utterance.

Methodology

The project involves the analysis of newly elicited production data as well as drawing on past research in the area as a basis for the construction of synthetic stimuli which incorporate either global or local voice source changes, as well as stimuli where both types of manipulations are incorporated.

The analytic study

Analysis is currently ongoing on recordings carried out in the semi-anechoic chamber of the Phonetics and Speech Laboratory at Trinity College Dublin. The subject is a young male speaker of Kerry Irish and a professional actor. He was chosen based on the fact that the current Kerry Irish male synthetic voice (available at www.abair.ie) is based on his voice, and this TTS system is envisaged as a potential testbed for the emerging model of affect manipulation. Declarative, semantically neutral Irish sentences with long open vowels were created, containing two or three (potentially) accented syllables. The subject was then asked to read the sentences so as to portray different affects, including *neutral*, *angry*, *sad*, *happy/excited*, *interested*, *relaxed/contented* and *bored*. Multiple repetitions were recorded, and the subject was advised to keep repetitions consistent in terms of intonation patterns.

From the recordings, a few exemplars were chosen on the basis of an informal listening session conducted at the Phonetics and Speech Laboratory involving the authors and colleagues. The selection of the two final exemplars was based on the criteria that (1) the utterances were perceived as good portrayals of a desired affect, and (2) the intonation pattern of the recording matched the expected intonation pattern which is typically found in the output of the TTS system, i.e., sentences had the typical falling contour of this dialect, and did not involve shifts in nuclear placement.

Analysis methods

The chosen exemplars were then manually inverse-filtered and the resulting estimate of the voice source signal was modelled using the Liljencrats-Fant (LF) model (Fant, Liljencrants and Lin, 1985). For details on the software system and the techniques used, see Ní Chasaide, Gobl and Monahan (1992) and Gobl and Ní Chasaide (2010). The LF model is a mathematical glottal flow model that allows extraction of voice source parameters. The main parameters of interest to this research are f0; EE (excitation strength) which is closely related to the overall intensity of the signal; and RD, a measure relating to the perceived tension in the voice, where RD is typically high for lax voice and low for tense/harsh voice (see Laver, 1980, for an in-depth description of voice quality). Samples were then segmented and annotated using the Praat software (Boersma and Weenink, 2020) for parameter visualization relative to the prosodic structure of the utterances (example shown in figure 1).

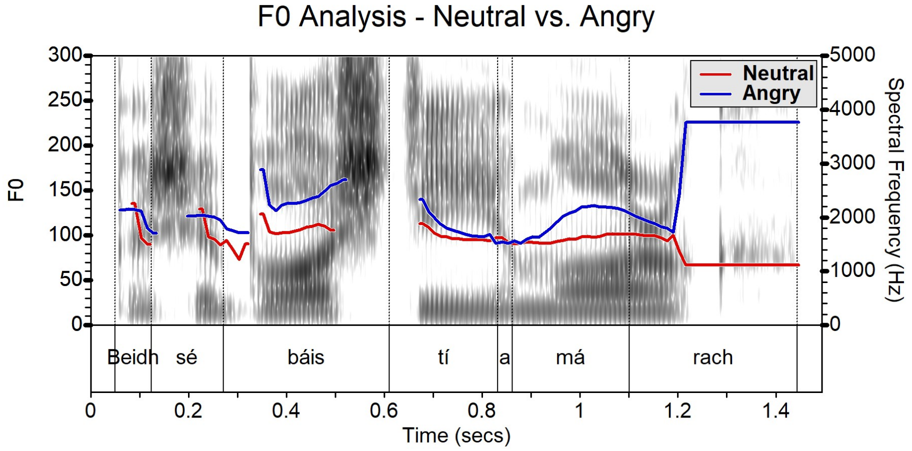

Figure 1. Preliminary visualization of the manual analysis comparing f0 measurements between *neutral* and *angry* exemplars for the sentence *Beidh sé báistí amárach* ("It will rain tomorrow").

Next steps

The construction of global and prosodically structured stimuli will be carried out by modifying the basic voice source characteristics of a synthesized utterance produced by the TTS voice for the speaker previously described using the system for voice analysis and synthesis GlórCáil (Murphy, Yanushevskaya, Ní Chasaide and Gobl, 2020). Recordings and resynthesis of a parallel set of stimuli will be carried out for a female synthetic voice for the Connemara dialect, whose basic intonation contours are roughly similar to the Kerry renditions. These resynthesized stimuli will then be used on perception tests in order to evaluate the effectiveness of the manipulations in the perception of affect by listener judges.

Acknowledgements

The research conducted in this publication was funded by the Irish Research Council under award number 17000, project number 211752.

References

Boersma, P., Weenink, D. 2020. Praat: doing phonetics by computer. Version 6.2.08.

Fant, G., Liljencrants, J., Lin, Q. 1985. A four-parameter model of glottal flow. STL-QPSR 26(4), 1-13, Speech, Music and Hearing, Royal Institute of Technology, Stockholm.

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

Gobl, C., Ní Chasaide, A. 2010. Voice source variation and its communicative functions. In Hardcastle, W. J., Laver, J., Gibbon, F. E. (eds.) 2010, The Handbook of Phonetic Sciences (Second Edition), 378-423. Oxford, Blackwell.

Juslin, P.N., Scherer, K.R. 2005. Vocal Expression of Affect. In Harrigan, H., Rosenthal, R. Scherer, K.R. (eds.) 2005, The New Handbook of Methods in Nonverbal Behaviour Research, 65-135. Oxford, Oxford University Press.

Laver, J. 1980. The phonetic description of voice quality. Cambridge, Cambridge University Press.

Murphy, A., Yanushevskaya, I., Ní Chasaide, A., Gobl, C. 2020. Testing the GlórCáil System in a Speaker and Affect Voice Transformation Task. 10th International Conference on Speech Prosody, 950-954. Tokyo, Japan.

Murphy, A., Yanushevskaya, I., Ní Chasaide, A., Gobl, C. 2022. Affect Expression: Global and Local Control of Voice Source Parameters. Speech Prosody 2022, 525-529. Lisbon, Portugal.

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.

Scherer, K.R. 2003. Vocal communication of emotion: A review of research paradigms. Speech Communication 40(1), 227-256.