Analysis of prosodic correlates of emotional speech data

Katarina Bartkova[1], Denis Jouvet[2]
[1]Université de Lorraine, CNRS, ATILF, F-54000 Nancy, France
[2]Universite de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

https://doi.org/10.36505/ExLing-2018/09/0004/000337

Abstract

The study of expressive speech styles remains an important topic as to their parameters detection or prediction in speech processing. In this paper, we analyze prosodic correlates for six emotion styles (anger, disgust, joy, fear, surprise and sadness), using data uttered by two speakers. The analysis is focused on the way pronunciations and prosodic parameters are modified in emotional speech, compared to neutral style. The analysis concerns speech pronunciation modifications, presence of pauses in sentences, and local prosodic behavior, with an emphasis set on the analysis of the prosody over prosodic groups and breathing groups.

Key words: expressive speech, emotions, prosodic groups, prosodic correlates.

Introduction

Prosody not only conveys information on the linguistic content of the vocal messages, but also on speaker's attitude and emotional state. In speech sciences, expressive speech is now attracting a lot of interest, as for example for speech synthesis (Schröder 2009) and for automatic recognition of emotions (Lanjewar 2013). For such studies, the collection of emotional speech data plays an important role. Several approaches have been envisaged for collecting emotional data; this includes the recording of natural emotions, the recording through induced situations, and the recording of acted speech by professional actors (Scherer 2003). This last mode is currently the most frequently used. Expressive speech synthesis has been developed for corpus-based (Iida 2003) and parametric-based approaches (Yamagishi 2004). As the recording of emotional data is difficult, several techniques have been investigated for adapting speech synthesis systems using small amount of emotional data (Inanoglu 2009) or for converting neutral speech into emotional speech (Tao 2006). It should be noted that parametric speech synthesis, needs to know the sequence of units, that is the sequence of sounds and pauses, for being able to produce the synthesized speech signal.

Following a previous study (Bartkova 2016) that has analyzed prosodic features on a global level, this paper focus on analyzing more local distributions associated to prosodic groups and breathing groups.

Corpus and features

The speech corpus used contains French sentences in six emotional styles (anger, disgust, joy, fear, sadness, and surprise), uttered by two professional speakers, one male and one female. Every emotional style contains about 50 sentences of various length. Each speaker has also uttered the same sentences in a neutral, reading style. This makes possible to study how pronunciations and prosodic parameters vary from neutral to emotional style.

For each sentence, the speech material has been aligned with its corresponding text using an automatic speech-text forced alignment procedure, and then manually checked and corrected if necessary. An automatic process has also been applied to localize prosodic boundaries, based on vowel duration, F0 slope, F0 delta, and pause occurrences,

Statistics on prosodic groups

Emotional speech is generally faster than neutral speech. Compared to neutral speech, the speaking is significantly higher (up to 28% faster) for fear and anger, slightly higher for disgust, joy and surprise, but significantly lower (about 20% slower) for sadness.

As the same sentences have been pronounced in neutral and emotional styles, it was possible to compare the presence and position of pauses. About 95% of the pauses observed in emotional data appears at the same place in the neutral pronunciation. However, except for sadness, 13% to 30% of the pauses observed in neutral pronunciation disappear in emotional data, which explains the increase in speaking rate for these emotional styles.

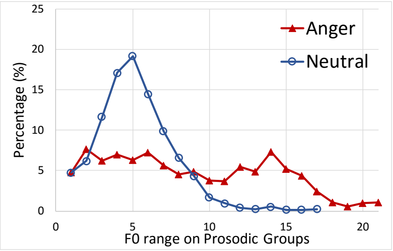

Figure 1: Histograms of F0 ranges over prosodic groups (average over the two speakers) in emotional and neutral speech styles.

Compared to neutral speech, the most noticeable differences in F0 range distributions are observed for anger and sadness (displayed in Figure 1). Larger F0 ranges are much more frequent for anger, and slightly more frequent for fear, surprise and joy (not represented in Figure 1 for lack of space); and smaller F0 ranges are more frequently observed for sadness.

Analysis of delta F0 at end of breathing groups

The histograms in Figure 2 display the distributions of the delta F0 measures at the end of the breathing groups (for lack of space, only two emotions are displayed). To better see the occurrences of falling and rising delta F0 values at the end of the breathing groups, the horizontal axis is symmetric (from -15 up to +15 semi-tones). Anger and fear styles comprise larger amounts of negative delta F0 values than the neutral style. The other noticeable difference is observed for sadness, displaying a sharp histogram of delta F0 values centered around zero; exhibiting rather flat F0 patterns. A tendency of using relatively flat or slightly falling F0 pattern is also observed in disgust.

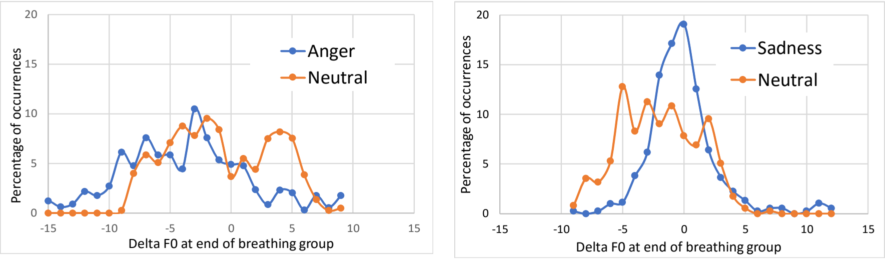

Figure 2: Histograms of delta F0 values at end of breathing groups (average over the two speakers) in emotional and neutral speech styles

Segmental level analysis

Sequences of phone segments associated to each word have been compared between emotion and neutral data to obtain statistics on the modification of the pronunciation of the words from neutral to emotional speech. Unlike other studies (Tahon 2016), few phoneme changes are observed in the data. The most common phoneme changes concern phonetic feature assimilations (mostly nasalization or voice feature assimilation) that are slightly more frequent in emotional speech than in neutral style. Some cases of liquid omissions are also observed, mainly in consonantal clusters. However, the main difference is mostly the omission of the schwa like vowel. In the emotional style a high number of schwa are omitted, and this vowel omission is place sensitive. In fact, the first and the last breathing groups of each record (a record corresponds to one or a few sentences) contain the highest number of schwa omissions (compared to neutral speech), 24% of the omissions are observed in the first breathing group and 27% in the last breathing group. Also, the number of schwa omissions is slightly emotion dependent, the highest percentage of schwa omissions is observed for disgust, fear and joy.

Conclusion

This paper has presented an analysis of prosodic correlates of emotional speech using a corpus of acted emotional speech. In comparison to neutral data, for the anger emotion, the speaking rate is higher, the number of pauses is significantly lower, larger F0 ranges over the prosodic groups are more frequent, as well as large negative delta F0 at end of breathing groups. Joy, disgust and fear emotions exhibit also higher speaking rate than neutral style. However, sadness exhibits a quite different behavior: compared to neutral speech, the speaking rate is lower, with a high number of inserted pauses. Unlike the other five emotional styles, in sadness style, there is a shift of the F0 range histogram towards lower values. As for the delta F0 at end of breathing groups, small values (rather flat F0 pattern) are more frequent.

Acknowledgments

This work has been partially supported by the CPER LCHN (Contrat Plan Etat Région "Langues, Connaissances et Humanités Numériques").

References

Schröder M. 2009. Expressive speech synthesis: Past, present, and possible futures. Affective information processing, 111-126. London: Springer.

Lanjewar R. B., Chaudhari D.S. 2013. Speech Emotion Recognition: A Review. Int. Journal of Innovative Technology and Exploring Engineering (IJITEE), 2.

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

Iida A., Campbell N., Higuchi F., Yasumura M. 2003. A corpus-based speech synthesis system with emotion. Speech Communication 40(1), 161-187.

Yamagishi J., Masuko T., Kobayashi T. 2004. HMM-based expressive speech synthesis-Towards TTS with arbitrary speaking styles and emotions. Proc. Special Workshop in Maui, Maui, Hawaı.

Inanoglu Z., Young S. 2009. Data-driven emotion conversion in spoken English. Speech Communication 51(3), 268-283.

Tao J., Kang Y., Li A. 2006. Prosody conversion from neutral speech to emotional speech. IEEE Trans. on Audio, Speech, and Language Processing 14(4), 1145-1154.

Bartkova K., Jouvet D., Delais-Roussarie E. 2016. Prosodic Parameters and Prosodic Structures of French Emotional Data. Speech Prosody 2016, Boston, United States.

Tahon M., Qader R., Lecorvé G., Lolive D. 2016. Optimal Feature Set and Minimal Training Size for Pronunciation Adaptation in TTS. SLSP'2016, 108-119.