EFFECT OF EMOTIONAL SPEECH ON ACOUSTIC SPEAKER DISCRIMINATION BASED ON THE RELATIVE SPEAKER CHARACTERISTIC

Juliane Höbel-Müller, Ronald Böck and Andreas Wendemuth

Chair of Cognitive Systems, Otto-von-Guericke-University Magdeburg, Germany
juliane.hoebel@ovgu.de

Introduction. Speaker discrimination (SD) has various applications such as speaker identification or speech turn segmentation. Unfortunately, SD robustness cannot only be impaired by noise and reverberation, but also by the intra-speaker variability of speech caused, for instance, by *affects* usually not considered during SD. Typically, acoustic SD is based on a speaker model, which is extracted directly from speaker's speech signal, which is free from noise and emotion, and its discriminative features. Investigating the influence of emotions on SD is in the focus of this study applying an approach, which is based on a relative characterization of the speaker, called Relative Speaker Characteristic (RSC) [1]. In [1] the performance of different classifiers using the RSC is analyzed regarding the effect of noise and the reduction of spectral bandwidth. This technique will be extended to emotions in the current paper.

Research Questions. In everyday life, speech is affected by different circumstances, in particular, various emotions. Current research on text-independent speaker identification shows that baseline systems mostly using Mel-frequency cepstral coefficients (MFCCs) can highly degraded by emotional speech (cf. [2]). For instance, speaker identification tests on emotional utterances from the benchmark corpus Berlin Database of Emotional Speech (EMO-DB) [3] results in Accuracy less than 60% (cf. [3]). As to our knowledge no work is done for RSC-based SD, the main focus of this study is on the influence of emotionally colored speech in SD. In this context, SD analyzed in checking whether two different speech signals are uttered by the same speaker or by two different speakers [4]. In this paper, the following question is answered: What is the RSC-based SD's performance with respect to emotional speech in comparison to emotionally neutral speech?

Dataset. To ensure a high quality of the emotional speech samples, EMO-DB [2] is used containing 494 samples. The database contains ten different sentences with neutral semantic content uttered by ten actors (five female) in the following seven basic emotions: anger, boredom, disgust, fear, joy, neutral, and sadness.

Experimental setup. The neutral utterances from EMO-DB served as training data for the binary SD classification in the entire experiment. A Support Vector Machine using the linear kernel was utilized. The "mfcc" feature set provided by openSMILE [5], comprising 13 MFCCs, their Delta and Delta-Delta values, was used to calculate the RSC acting as a meta-feature for SD. Additionally, as the RSC is a non-standardized measure, standardization was utilized. Regarding the conducted experiments, the classifier's training is done using a balanced distribution of samples, which are based only on emotionally neutral utterances. A stratified 10-fold cross-validation was conducted. In the baseline experiment, SD was evaluated using the neutral emotion, only. By evaluating the influence of each emotion on SD, we answered our research question.

Results. We evaluated SD's performance calculating the F1 Score. When emotionally neutral speech is used for both training and testing, the F1 Score is 91.75%. Using all emotions for testing, except the neutral one, the F1 Score is 82%. Moreover, by obtaining a Correct Discrimination Rate of 72%, state-of-the-art systems using GMMs and MFCCs (cf. [3]) can be outperformed. SD is influenced differently by each emotion. The best F1 Scores were obtained for boredom (86.53%), joy (84.1%), anger (82.52%), and fear (80.8%). Sadness (68%) and disgust (73%) represent more difficult emotional states for SD.

Conclusion and Outlook. When emotions effect the human voice, the performance of RSC-based SD decreases. Especially, discriminating speakers whose voices are being altered by unpleasant and mild emotions is difficult. However, it can be noted positively that most of the basis emotions (boredom, joy, anger, and fear) only have a small impact on the RSC-based SD. EMO-DB contains prototypical expressions of full-blown emotions, which are rarely found in everyday life. Therefore, SD analysis could be continued on vocalizations of more subtle affects.

References

[1] OUAMOUR, S.; GUERTI, M.; SAYOUD, H. A NEW RELATIVISTIC VISION IN SPEAKER DISCRIMINATION. *CANADIAN ACOUSTICS*, 2008, 36. JG., NR. 4, S. 24-35.

[2] GHIURCAU, MARIUS VASILE; RUSU, CORNELIU; ASTOLA, JAAKKO. A STUDY OF THE EFFECT OF EMOTIONAL STATE UPON TEXT-INDEPENDENT SPEAKER IDENTIFICATION. IN: *PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP), 2011*. S. 4944-4947.

[3] BURKHARDT, FELIX, ET AL. A DATABASE OF GERMAN EMOTIONAL SPEECH. IN: *INTERSPEECH*. 2005. S. 1517-1520.

[4] ROSE, PHIL, ET AL. FORENSIC SPEAKER DISCRIMINATION WITH AUSTRALIAN ENGLISH VOWEL ACOUSTICS. *ICPHS XVI SAARBRUCKEN*, 2007, 6. JG., NR. 10.

[5] EYBEN, FLORIAN; WÖLLMER, MARTIN; SCHULLER, BJÖRN. OPENSMILE: THE MUNICH VERSATILE AND FAST OPEN-SOURCE AUDIO FEATURE EXTRACTOR. IN: *PROCEEDINGS OF THE 18TH ACM INTERNATIONAL CONFERENCE ON MULTIMEDIA*. ACM, 2010. S. 1459-1462.