EMOTION RECOGNITION FROM DISTURBED SPEECH - TOWARDS AFFECTIVE COMPUTING IN REAL-WORLD IN-CAR ENVIRONMENTS

Alicia Flores Lotz[1], Fabian Faller[2], Ingo Siegert[1], Andreas Wendemuth[1]

[1] Institute for Information and Communications Engineering, Cognitive Systems Group, Otto-von-Guericke University, 39016 Magdeburg, Germany

[2] Interior Systems & Technology Advanced Development, Continental Automotive GmbH, 64832 Babenhausen, Germany

Introduction It is well known that speech emotion recognition rates are high (> 90%) for acted emotions under clear recording conditions. Further, research has shown to date that these rates drop considerably (< 60%) if naturalistic emotions are considered [6], and if data is disturbed with artificial noise or if recordings are performed in noisy environments [2]. A less investigated scenario that we will discuss are in-car settings. Here, particular naturalistic emotions will be observed, and the acoustic recordings will be naturally perturbed by convoluted in-car noise which is of a special nature. Previous investigations have only superimposed clean speech to different car noise types and noise levels. We will give an overview on differences in quality and performance of classification tasks conducted in this environment. The full paper will also motivate the importance of why emotion recognition in in-car environment is necessary for the assessment of typical driving situations, and how emotion recognition results can be used by the in-car systems to enhance driving safety and comfort.

Database To achieve controlled test conditions, a driving simulator of Continental Automotive GmbH was used. To obtain comparability with published results under different conditions, the data samples of the Vera am Mittag corpus [4] and the Berlin Database of Emotional Speech [1] were used. By replaying the samples in the simulator, the simulated car noise was not superimposed but convoluted with the speech samples. These databases cover both, categorical, high quality, expressive, acted emotions and dimensional, low quality, spontaneous, scripted emotions. In total, 3 hours of acoustic material was recorded.

Recording Setup The recordings of the replayed emotional samples were conducted using two directional shotgun microphones placed at the A-pillars of the simulator vehicle. The samples were played back from loudspeakers mounted at head heights on the drivers seat. Two different recording scenarios were used to obtain samples only influenced by the in-car acoustics (simulator turned off, no environmental noise present) and disturbed with simulated highway noise (simulator turned on, driving autonomously).

Methodology To evaluate the quality decrease of the disturbed speech samples, the signal-to-noise ratio was computed. Additionally, a quality measure based on the Compression Error Rate (CER) presented in [5] was considered. To also evaluate the influence of noise in speech emotion recognition, state-of-the-art classification experiments were carried out using Support Vector Machines (SVM) and the "'emobase"' feature set [3]. The significance of the results was then analyzed using the statistical evaluation methods ANOVA and t-test.

Results The full paper will present the results of the investigations described above and discusses the necessity of adapting emotion recognizers for different in-car convolutive noise levels. A clear statement on the performance reduction of classifiers trained on clean in contrast to disturbed speech in cars will be given.

References

[1] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A database of german emotional speech. In: Proc. of the 9th International Conference on Spoken Language Processing, INTERSPEECH 2005. pp. 1517–1520. Lisbon, Portugal (2005)

[2] Chenchah, F., Lachiri, Z.: Speech emotion recognition in noisy environment. In: Proc. of the 2nd International Conference on Advanced Technologies for Signal and Image Processing, ATSIP 2016. pp. 788–792. Monastir, Tunisia (2016)

[3] Eyben, F., Wöllmer, M., Schuller, B.: openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor. In: Proc. of the ACM MM-2010. p. s.p. Firenze, Italy (2010)

[4] Grimm, M., Kroschel, K., Narayanan, S.: The vera am mittag german audio-visual emotional speech database. In: Proc. of the 2008 IEEE ICME. pp. 865–868. Hannover, Germany (2008)

[5] Lotz, A.F., Siegert, I., Maruschke, M., Wendemuth, A.: Audio compression and its impact on emotion recognition in affective computing. In: Trouvain, J., Steiner, I., Möbius, B. (eds.) Elektronische Sprachsignalverarbeitung 2017. Studientexte zur Sprachkommunikation, vol. 86, pp. 1–8. TUDpress (2017)

[6] Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., Wendemuth, A.: Acoustic Emotion Recognition: A Benchmark Comparison of Performances. In: Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2009. pp. 552–557. Merano, Italy (2009)