ACOUSTIC DETECTION OF CONSECUTIVE STAGES OF SPOKEN INTERACTION BASED ON SPEAKER-GROUP SPECIFIC FEATURES

Ronald Böck, Olga Egorow, Andreas Wendemuth
Cognitive Systems Group, Otto von Guericke University, 39016 Magdeburg, Germany
{first_name.family_name}@ovgu.de

Introduction. Naturalistic human communication is affected by various circumstances. Prominent effects are the current situation, the communication interlocutors as well as personal characteristics like sex and age [1]. Most aspects of human-human interaction also occur in human-computer interaction (HCI) [2]. Moreover, it can be assumed that each interaction is influenced by various stages of a communication [3]. On one hand, they make the interaction more challenging to process automatically, but on the other hand, they can be seen as an additional source of information. In this study we explored the detection of consecutive stages based on specific features investigated in [4].

Research Questions. The aim of this study is to detect different parts of HCI, in particular, two consecutive stages in a spoken interaction with a companion-like technical system. For this, we extended the acoustic analyses presented in [4] to all participants in the LAST MINUTE Corpus and derived minimal feature sets suitable for stage detection. This is done in a naive, naturalistic communication, where the subject is (re-)acting in a non-scripted way. In particular, we are interested in two research questions: Can spectral or prosodic features be analytically identified across speakers indicating differences in consecutive interaction stages across speakers to form minimal feature sets? Can these feature sets be used to detect interaction stages?

Data Set. For this study, we used the LAST MINUTE Corpus (LMC), consisting of naturalistic HCI recordings. The recorded interactions are divided in four distinct stages representing different situations a user can face [5]. Each of these stages is marked by a so-called barrier [3] that allows to align the users' utterances with a certain situation. In the current study, we analysed the utterances of all 89 participants who contributed acoustically to LMC (48 female, 41 male) in two sub-scenarios: a more relaxed stage and a more challenging stage.

Experimental Setup. In general, we implemented the setup as presented in [4]: 52 spectral and prosodic features were extracted using openSMILE's "emobase" configuration [6], resulting in one mean value for each feature and utterance which were averaged over all utterances per stage. Statistical analyses provided significant differences for certain features compared inter-stage-wise. These features were collected to form feature sets to be used in the detection experiments. Besides the two identified feature sets (1) containing only highly significant features and 2) containing all significant features) we test the "emobase" collection as well as a set suitable for addressee detection [7]. For detection, Support Vector Machines (SVM) with linear and polynomial kernel as well as Random Forest were applied using Weka [8].

Results. In the current study, we investigated the acoustic differences in consecutive interaction stages of 89 participants of LMC. Based on statistical methods, we identified two minimal feature sets containing either 14 highly significant or 29 significant features (for detection, we added finally all statistic of the identified features; in total 409 features). Most prominent are MFCC- and LSP-related features. In the detection experiments, we compared our feature sets to "emobase" (988 features) [6] and a speaker addressee detection set (700 features) [7]. The table presents the results using Random Forest. For the 14 features' set we achieved an unweighted average recall (UAR) of 0.591 (±0.109) applying an SVM. It is to be noticed that we work with naturalistic material with low expressiveness resulting in lower UAR values.

Feature Set	14 Features	Extended 29 Features	Speaker Addressee	Emobase
Unweighted Avg. Recall (Random Forest)	0.564 (±0.104)	0.645(±0.101)	0.656(±0.102)	0.654(±0.111)

Discussion and Conclusion. Given the results presented in the table, a small feature set, especially feasible for HCI systems under mobile conditions, can indicate different interaction stages or possible changes. Afterwards, a more complex feature set is used for detection of particular stages. In our experiments, we found that the extended 29 features' set (409 features) performs in the same range as sets with 700 or 889 features ("emobase") which is also meaningful for restricted mobile devices.

In the present paper, we could show that with a standardised feature set, interaction stages of 89 participants of LMC corpus can be identified and further, we identified a minimal feature set for the detection of consecutive interaction stages in naturalistic HCI based on statistical analyses.

References:

[1] Siegert, I.; D. Philippou-Hübner; K. Hartmann; R. Böck & A. Wendemuth (2014). 'Investigation of Speaker Group-Dependent Modelling for Recognition of Affective States from Speech'. Cognitive Computation 6(4):892–913.

[2] Nass, C.; J. Steuer & E.R. Tauber (1994). 'Computers are social actors'. In: Proceedings of the SIGCHI conference on Human factors in computing systems. Boston, USA:ACM, pp.72–78.

[3] Frommer, J.; D. Rösner; M. Haase; J. Lange; R. Friesen & M. Otto (2012). Detection and Avoidance of Failures in Dialogues – Wizard of Oz Experiment Operator's Manual. Pabst Science Publishers.

[4] Böck, R.; Egorow, O. & A. Wendemuth (2017). 'Speaker-Group Specific Acoustic Differences in Consecutive Stages of Spoken Interaction'. In: Proceedings of the 27. Konferenz Elektronische Sprachsignalverarbeitung. Saarbrücken, Germany: TUD Press, pp. 211-218.

[5] Rösner, D.; J. Frommer; R. Andrich; R. Friesen; M. Haase; M. Kunze; J. Lange & M. Otto (2012). 'LAST MINUTE: a Novel Corpus to Support Emotion, Sentiment and Social Signal Processing'. In: Proceedings of the Eight International Conference on Language Resources and Evaluation. Istanbul, Turkey: ELRA, pp. 82–89.

[6] Eyben, F.; M. Wöllmer & B. Schuller (2009). 'OpenEAR - Introducing the munich open-source emotion and affect recognition toolkit'. In: Proceedings of 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops. Amsterdam, The Netherlands: IEEE, pp. 576–581.

[7] Tang, S. (2017). 'Analysis of acoustic features and automatic recognition experiments for conversation addressee detection'. Master Thesis. Otto von Guericke University Magdeburg, Germany.

[8] Hall, M.; Eibe, F.; Holmes, G.; Pfahringer, B.; Reutemann, P. & I.H.Witten (2009) 'The WEKA Data Mining Software: An Update'. SIGKDD Explor. Newsl. 11:10–18.