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
ronald.boeck@ovgu.de

Abstract: Spoken human interaction comprises a variety of tasks and behaviours. Since the state of a human is affected by various circumstances, it can change during an interaction – for example, depending on the task at hand. Such interaction stages occur in human-human interaction as well as in human-computer interaction, and therefore are object of automatic detection and classification. In this paper, we investigate several feature sets with respect to their suitability to the task of interaction stages classification.

1 Introduction

A natural human conversation can include several distinct stages with their conversational tasks and certain behavioural and interactional patterns, influenced by the current situation, the interlocutors and their roles and personal characteristics like sex and age [1]. These influences not only play a role in human-human interaction (HHI), they are also important for human-computer interaction (HCI) [2]. On one hand, the differences elicited by such factors render conversations unique and therefore are hard to process automatically for the purposes of HCI. But on the other hand, they also offer additional information that can be used for designing better interaction systems that are able to adapt, depending on the needs of the user, in an anticipative way.

In order to use this additional information on the user's current state, we need to assess it. In a previous study, we preliminarily explored the influence of sex and age on certain acoustic features of speech in different stages of close-to-real-life HCI [3], which is now being extended. In the present study, we aim to investigate the suitability of the previously explored features for the task of the detection of such consecutive interaction stages (cf. section 4) in the same close-to-real-life HCI. For this, we compared the performance of the generated feature sets with the benchmark set *emobase* [4] and a set feasible for addressee detection [5].

2 Research Questions

The aim of this study is to detect different parts of HCI, in particular, consecutive stages in a spoken interaction with a companion-like technical system. In particular, two stages are observed which are positioned consecutively in the course of an interaction but are separated by a so-called barrier (cf. section 4 and Table 1). For this, we extended the acoustic analyses presented in [3] to all participants in the LAST MINUTE Corpus (LMC) 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 but based on a well-established study design (cf. [6]).

In particular, we are interested in two research questions:

Q1: Can spectral or prosodic features indicating differences in consecutive interaction stages be analytically identified to form minimal feature sets?

Q2: Can these feature sets be used to discriminate interaction stages?

3 Related Work

In [7], it is discussed that the detection and classification of an affective user state is an important issue in HCI. Especially, in a close-to-real-life interaction where the user is not limited in wording, behaviour, and affect, this is a challenging task (cf. [8]). Aiming for the handling of such situations, the focus in the spoken communication community shifted from acted emotions, as provided for example in the well-received Berlin Database of Emotional Speech [9] to more natural emotions and dispositions (cf. [10]) or so-called "in the wild" scenarios (cf. [11]).

Various investigations concerning automatic emotion recognition have been reported (cf. e.g. [12, 13]) and have been extended to dispositions recognition (cf. [14]), also several analyses are presented on the LMC (cf. e.g. [1, 15, 16]) that was also used in the current study (cf. section 4). Considering the literature, it can be stated that acoustic emotion recognition using machine learning methods still does not provide satisfying results for "in the wild" scenarios. Besides the optimisation of classification techniques, a crucial issue is the question which features best describe the emotional content of spoken communication as well as possible changes of emotions (cf. e.g. [17]). Currently, various feature sets are being recommended, most prominent the *Geneva Minimalistic Acoustic Parameter Set* (GeMAPS) [18] and versions of the *emobase* feature set (cf. e.g. [4]). Analytical results for spectral features applied on emotional states based on distinct emotions are reported (cf. e.g. [19, 20]). To the best of our knowledge, the only comparison of spectral and prosodic features changes and how exactly they change in different stages of close-to-real-life HCI is presented in [3], which provides the basis to the current investigation.

In search of suitable feature sets, we considered related tasks, such as speaker identification and addressee detection. We assume that the distinguishing of two addressees, either a human or a technical system, is comparable to the issue of separating two interaction stages.
Automatic addressee detection via audio has already been in focus for some time, for example as part of the INTERSPEECH 2017 computational paralinguistics challenge – here the task was to detect whether the addressee was a child or an adult. For this task, the organisers of the challenge offered two different feature sets. The first set is the ComParE acoustic feature set containing 6737 static features resulting from various functionals of 65 Low-Level-Descriptor (LLD) contours. This feature set is explained in detail in [21, 22]. The second set consists of bag-of-audio-words features, where the audio chunks are represented as histograms of LLDs. More details on this feature set can be found in [23].
For distinguishing between a human and a technical addressee, a sophisticated mix of lexical and acoustic-prosodic features is used in [24]: besides lexical features obtained from automatic speech recognition, the authors compute energy contour features, voice quality features, spectral tilt features, and delta energy at voicing onsets and offsets. Statistical analyses of features conducted in a naturalistic addressee detection scenario are obtained in [5] proposing a suitable feature set.

4 Data Set

For this study, we used the LAST MINUTE Corpus (LMC) [25], consisting of naturalistic HCI recordings of close-to-real-life HCI using a Wizard-of-Oz (WOZ) setup. In the experiments,

Interaction Stage	Triggering Event	Task
Baseline	–	Introduction to the system
Listing	50% of categories finished	Choosing items from a list
Challenge	Reaching weight limit	Deleting items
Waiuku	Revealing destination	Re-organising suitcase

Table 1 – Overview of the interaction stages during the LMC experiments and their respective barriers and tasks.

the participants have to accomplish a suitcase packing task while the conditions grow ever more complicated. The recorded interactions are divided into four distinct stages (cf. Table 1) representing situations with an increasing difficulty [25]. Each of these stages is marked by a so-called barrier [6] that allows to align the users' utterances with a certain situation. The interactions are built around an imaginary trip to an unknown location that shall be prepared in limited time. In the first interaction stage, the participants get to know the system and introduce themselves in order to get comfortable. In the second stage, the participants have the task to pack a suitcase choosing items from a list via voice commands. In the next stage, they get the information that the weight limit is reached and have to delete some items. In the last stage, they receive the information that the trip is a winter trip instead of a summer trip, which leads to a complete re-organisation of the suitcase. In conclusion, the participants are asked on their overall interaction experience. An overview of the interaction stages is given in Table 1.

In the current study, we analysed the utterances of 89 participants, providing feasible audio quality in the LMC (48 female, 41 male, 43 younger than 30 years, 46 older than 60 years), in two sub-scenarios: the more relaxed stage "listing" and the more challenging stage "challenge". Both consecutive stages are separated by the "weight limit barrier" indicating that the airline specified luggage weight is reached, resulting in a participant-initiated restructuring process including deletion and choice of particular items.

5 Experimental Setup

In general, we implemented the setup as presented in [3]: 52 spectral and prosodic features, such as intensity, Mel-Frequency Cepstral Coefficients (MFCCs), Line Spectral Pairs (LSPs), etc., were extracted using openSMILE's *emobase* configuration [4], representing one mean value for each LLD. Given the *emobase* setting, one mean value is obtained for each utterance. These values were then averaged over all utterances of a stage per speaker. Based on the calculated standard deviation for each mean LLD across all speakers, analyses provided remarkable differences for certain features compared inter-stage-wise. These features formed feature sets to be used in the detection experiments.

Besides the two identified feature sets (cf. section 6.1) 1) containing *only* highly remarkable features (the calculated number of changes across speakers $k \geq 2.5\sigma$) and 2) containing *all* remarkable features ($k \geq \sigma$), we tested the *emobase* feature set (serving as baseline) as well as a feature set suitable for addressee detection [5]. The latter set is inspired by the idea of knowledge transfer: In addressee detection, the aim is to discriminate utterances from two situations, namely those spoken to a human and those approaching a technical device, since this task resembles the distinguishing between two interaction stages that we investigated.
Conducting the classification process, a Leave-One-Speaker-Out (LOSO) evaluation was applied where the material was 1) used as it is and 2) standardised separately for training and testing (cf. [26]). Since we were mainly interested in the potential classification power of each feature set, no particular fine-tuning of the classifiers was performed. For detection, the distance-based Support Vector Machine (SVM) with linear and polynomial kernel as well as the

non-distance-based Random Forest (RF) were applied using Weka [27] reporting Unweighted Average Recall (UAR) values.

6 Analyses of Consecutive Interaction Stages

6.1 Feature Sets

In the present study, we investigated the acoustic differences in consecutive interaction stages (cf. research question Q1) of 89 participants of the LMC. Based on statistical methods explained in section 5 and [3], we identified various features with are remarkably different in the consecutive stages across speakers. It is to be noticed that each difference can be directed towards an increasing or a decreasing of the particular values. Most prominent are MFCC- and LSP-related features in both the absolute values and respective derivatives (cf. Table 2). This finding highlights also the discriminative power of spectral features distinguishing interaction stages (for emotions from speech cf. e.g. [17, 18]). Interestingly, prosodic features are less represented in the identified ones. This might be influenced by the close-to-real-life conditions of the recordings (cf. section 4) where the expressiveness is generally lower related to smaller differences in prosodic features.

Table 2 – List of features showing remarkable differences in consecutive stage on LMC. Highly remarkable features are highlighted. Int refers to intensity, Loud refers to loudness, and Δ indicates the particular delta values.

	Low Level Descriptors									
LLD Instances	Int	Loud	MFCC1	LSP0	PCM_ZCR	F0	ΔInt	ΔLoud	ΔMFCC1	ΔLSP1
			MFCC2	LSP1					ΔMFCC2	ΔLSP2
			MFCC3	LSP2					ΔMFCC9	ΔLSP3
			MFCC4	LSP3					ΔMFCC12	ΔLSP4
			MFCC6	LSP4						ΔLSP6
			MFCC12	LSP5						ΔLSP7
				LSP6						

Based on the statistical analyses, we created two minimal feature sets (cf. Table 2) which will be used for discrimination of consecutive interaction stages in section 6.2. We identified 14 highly remarkable features, mainly represented by MFCCs and LSPs that are highlighted in Table 2. Besides the spectral features intensity, PCM_ZCR, Δ intensity, and Δ loudness constitute the so-called F_{14} set. Regarding *all* 29 remarkable features the so-called F_{29} set can be constructed (referring to all features listed in Table 2). Based on F_{29}, we constructed an extended feature set including LLDs' functionals. This results in 409 features, the so-called F_{409} set. We chose this option, inspired by *GeMAPS* (cf. [18]), achieving an intermediate number of features compared to *emobase*. All three sets are applied in the detection experiments.

6.2 Classification Results

In the detection experiments (cf. research question Q2), distinguishing between two consecutive stages, we compared our feature sets identified by statistical analysis to the *emobase* set F_{EB} (988 features) [4] and a set feasible for addressee detection F_{AD} (700 features) in a close-to-real-life setup [5]. Furthermore, as discussed in section 6.1, we used F_{409} as possibility to incorporate functionals like in *GeMAPS* or *emobase*. As stated in section 5, two classification techniques have been considered, namely SVM and RF.

The classification results in terms of UAR applying an SVM with linear kernel are shown in Figure 1. For the F_{14} set, we achieved an UAR of 0.591 ($\pm$0.109). This performance is an improvement of $\sim$5% relative to the baseline (*emobase*). It is to be noticed that we work with naturalistic material with low expressiveness resulting in lower UAR values. Given the variance (cf. Figure 1), it can be stated that the results utilising different features sets do not vary significantly.

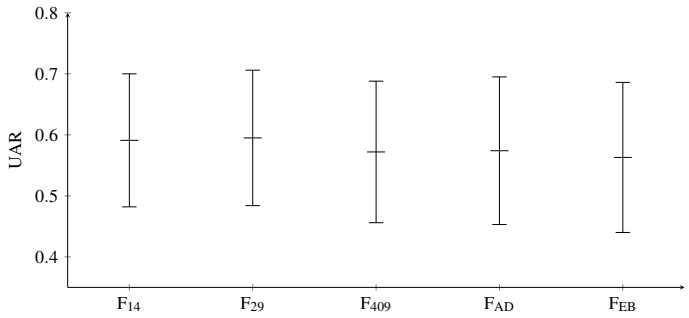

Figure 1 – Mean values of unweighted average recall (UAR) and variances of respective feature sets applying a linear Support Vector Machine on standardised data.

The detection results applying SVMs do not show any statistical significance. To eliminate any influence of the classifier, we also conducted detection experiments utilising the non-distance-based RF classifiers. The results are presented in Table 3. Comparing the results of SVM and RF, we see a clear difference in the performances of the particular classifier approaches related to the feature sets. Especially, the SVMs benefit from the statistically identified features in F_{14} and F_{29}. In contrast, RF techniques already internally provide a feature selection (cf. [28]) which benefits from a larger number of features. Therefore, it can be assumed that smaller, already well-defined feature sets reduce the selection power of the RF. However, the internal selection process (cf. [28]) further reduces the number of utilised features, resulting in a decrease of UAR values (cf. Table 3).

Table 3 – Unweighted average recall (UAR) and variance for Random Forest classifiers on five different feature sets.

Feature Set	$\mathbf{F}_{14}$	$\mathbf{F}_{29}$	$\mathbf{F}_{409}$	$\mathbf{F}_{AD}$	$\mathbf{F}_{EB}$
UAR	0.564 ($\pm$0.104)	0.578 ($\pm$0.108)	0.645 ($\pm$0.101)	0.656 ($\pm$0.102)	0.654 ($\pm$0.111)

7 Discussion

Given our approach just LLDs were considered in the analytical evaluation of features to discriminate interaction stages. The averaging, reflected by the mean values per feature, is used to generate a measure on utterance level. This approach is reasonable since it can be assumed that small differences per frame can be neglected to obtain a broader understanding of distinguishing characteristics. Besides identifying remarkable features, we conducted detection experiments for consecutive interaction stages based on the LMC. For this, distance- and non-distance-based classifiers were applied to avoid any influence caused by internal classification methods.

Given the results presented in Figure 1 and Table 3, we can state that the proposed feature

sets do not differ significantly in the detection performance. The distance-based SVM on F_{14} achieved a $\sim$5% better UAR compared to the baseline. Taking functionals into account, the SVMs' performance decreased slightly which may be a result of the higher complexity to find appropriate separations in the larger feature space. In contrast, using a non-distance-based method like RF, functionals help to distinguish interaction stages as the internal RF's feature selection benefits from a larger number of features. Nevertheless, the F_{409} set performed almost equally compared to larger feature sets containing 700 or 988 features (cf. Table 3). The idea of transferring features suitable for addressee detection (cf. [5]) to interaction stage detection (cf. section 5) did not improve the classification performance compared to F_{409} or *emobase*.
Since the different feature sets perform similarly in the given task (no statistical differences), already a small feature set, especially feasible for HCI systems under mobile conditions, can indicate different interaction stages. Even a more complex feature set comprising 409 features halved the values to be considered compared to the common 988 features in *emobase*. This is an aspect especially important for mobile devices with limited resources.

8 Conclusion

In this paper, we analysed interaction stages of 89 participants of the LMC [25], a naturalistic HCI corpus. In particular, we investigated two research aspects regarding suitable features for interaction stage discrimination and classification performances based on the LMC.

Based on statistical analyses (cf. section 5 and [3]), we identified remarkable features which provide discriminative power for interaction stages. As given in Table 2, most features are related to MFCCs and LSPs. Therefore, we contributed to the still ongoing discussion on the question which features are feasible for recognition tasks from speech.

Additionally, we conducted various detection experiments based on the identified features compared to two "baseline" sets, namely an addressee detection feature set [5] and *emobase* [4]. For both classification approaches no significant differences in the performance of the particular sets were seen. Therefore, we can conclude that our small feature sets provided a similar recognition performance. In particular, an UAR of 65% with a RF-based approach and utilising 409 features was achieved (cf. Table 3). This is of certain interest especially for devices with limited resources, for example for mobile devices.

Acknowledgement

We acknowledge support by the project "Intention-based Anticipatory Interactive Systems" (IAIS) funded by the Federal State of Sachsen-Anhalt, Germany. Further, we thank the projects "Mova3D" (grant number: 03ZZ0431H) and "Mod3D" (grant number: 03ZZ0414) funded by 3Dsensation within the Zwanzig20 funding program by the German Federal Ministry of Education and Research (BMBF).

References

[1] SIEGERT, I., D. PHILIPPOU-HÜBNER, K. HARTMANN, R. BÖCK, and A. WENDEMUTH: *Investigation of Speaker Group-Dependent Modelling for Recognition of Affective States from Speech*. *Cognitive Computation*, 6(4), pp. 892–913, 2014.

[2] NASS, C., J. STEUER, and E. R. TAUBER: *Computers Are Social Actors*. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 72–78. ACM, Boston, USA, 1994.

[3] BÖCK, R., O. EGOROW, and A. WENDEMUTH: *Speaker-Group Specific Acoustic Differences in Consecutive Stages of Spoken Interaction*. In *Proceedings of the 28. Konferenz Elektronische Sprachsignalverarbeitung*, pp. 211–218. TUD Press, Saarbrücken, Germany, 2017.

[4] EYBEN, F., F. WENINGER, F. GROSS, and B. SCHULLER: *Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor*. In *Proceedings of the 21st ACM International Conference on Multimedia*, pp. 835–838. ACM, Barcelona, Spain, 2013.

[5] TANG, S.: *Analysis of acoustic features and automatic recognition experiments for conversation addressee detection*. Master's thesis, Otto von Guericke University Magdeburg, 2017.

[6] FROMMER, J., D. RÖSNER, M. HAASE, J. LANGE, R. FRIESEN, and M. OTTO: *Detection and Avoidance of Failures in Dialogues – Wizard of Oz Experiment Operator's Manual*. Pabst Science Publishers, 2012.

[7] PICARD, R. W.: *Affective Computing for HCI*. In *Proceedings of the International Conference on Human-Computer Interaction*, pp. 829–833. Lawrence Erlbaum, Munich, Germany, 1999.

[8] WARD, R. D. and P. H. MARSDEN: *Affective Computing: Problems, Reactions and Intentions*. *Interacting with Computers*, 16(4), pp. 707–713, 2004.

[9] BURKHARDT, F., A. PAESCHKE, M. ROLFES, W. SENDLMEIER, and B. WEISS: *A Database of German Emotional Speech*. In *Proceedings of the INTERSPEECH-2005*, pp. 1517–1520. ISCA, Lisbon, Portugal, 2005.

[10] BÖCK, R.: *Multimodal Automatic User Disposition Recognition in Human-Machine Interaction*. Ph.D. thesis, Otto von Guericke University Magdeburg, 2013.

[11] GRIFFITHS, P. E. and A. SCARANTINO: *Emotions in the wild: The situated perspective on emotion*. In *The Cambridge handbook of situated cognition*, pp. 437–453. Cambridge University Press, 2009.

[12] RINGEVAL, F., S. AMIRIPARIAN, F. EYBEN, K. SCHERER, and B. SCHULLER: *Emotion Recognition in the Wild: Incorporating Voice and Lip Activity in Multimodal Decision-Level Fusion*. In *Proceedings of the 16th ICMI*, pp. 473–480. ACM, Istanbul, Turkey, 2014.

[13] SIKKA, K., K. DYKSTRA, S. SATHYANARAYANA, G. LITTLEWORT, and M. BARTLETT: *Multiple kernel learning for emotion recognition in the wild*. In *Proceedings of the 15th ICMI*, pp. 517–524. ACM, Sydney, Australia, 2013.

[14] BIUNDO, S. and A. WENDEMUTH (eds.): *Companion Technology - A Paradigm Shift in Human-Technology Interaction*. Springer, 2017.

[15] FROMMER, J., B. MICHAELIS, D. RÖSNER, A. WENDEMUTH, R. FRIESEN, M. HAASE, M. KUNZE, R. ANDRICH, J. LANGE, A. PANNING, and I. SIEGERT: *Towards Emotion and Affect Detection in the Multimodal LAST MINUTE Corpus*. In *Proceedings of the 8th LREC*, pp. 3064–3069. ELRA, Istanbul, Turkey, 2012.

[16] EGOROW, O. and A. WENDEMUTH: *Detection of Challenging Dialogue Stages Using Acoustic Signals and Biosignals*. In *Proceedings of the WSCG 2016*, pp. 137–143. ACM, Plzen, Czech Republic, 2016.

[17] TAHON, M. and L. DEVILLERS: *Towards a small set of robust acoustic features for emotion recognition: Challenges. IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(1), pp. 16–28, 2016.

[18] EYBEN, F., K. R. SCHERER, B. W. SCHULLER, J. SUNDBERG, E. ANDRÉ, C. BUSSO, L. Y. DEVILLERS, J. EPPS, P. LAUKKA, S. S. NARAYANAN ET AL.: *The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing*. *IEEE Transactions on Affective Computing*, 7(2), pp. 190–202, 2016.

[19] KIENAST, M. and W. F. SENDLMEIER: *Acoustical analysis of spectral and temporal changes in emotional speech*. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*. ISCA, Newcastle, UK, 2000.

[20] YILDIRIM, S., M. BULUT, C. M. LEE, A. KAZEMZADEH, Z. DENG, S. LEE, S. NARAYANAN, and C. BUSSO: *An acoustic study of emotions expressed in speech*. In *Proceedings of the 8th International Conference on Spoken Language Processing*. ISCA, Jeju Island, Korea, 2004.

[21] HANTKE, S., F. WENINGER, R. KURLE, F. RINGEVAL, A. BATLINER, A. E.-D. MOUSA, and B. SCHULLER: *I hear you eat and speak: Automatic Recognition of Eating Condition and food type, use-cases, and impact on ASR performance*. *PloS one*, 11(5), 2016.

[22] EYBEN, F.: *Real-time speech and music classification by large audio feature space extraction*. Springer, 2015.

[23] SCHMITT, M. and B. SCHULLER: *openXBOW – Introducing the Passau open-source crossmodal bag-of-words toolkit*. *Journal of Machine Learning Research*, 18(96), pp. 1–5, 2017.

[24] SHRIBERG, E., A. STOLCKE, and S. V. RAVURI: *Addressee detection for dialog systems using temporal and spectral dimensions of speaking style.* In *Proceedings of the INTERSPEECH-2013*, pp. 2559–2563. ISCA, Lyon, France, 2013.

[25] RÖSNER, D., J. FROMMER, R. ANDRICH, R. FRIESEN, M. HAASE, M. KUNZE, J. LANGE, and M. OTTO: *LAST MINUTE: a Novel Corpus to Support Emotion, Sentiment and Social Signal Processing*. In *Proceedings of the 8th LREC*, pp. 82–89. ELRA, Istanbul, Turkey, 2012.

[26] BÖCK, R., O. EGOROW, I. SIEGERT, and A. WENDEMUTH: *Comparative Study on Normalisation in Emotion Recognition from Speech.* In *Intelligent Human Computer Interaction*, no. 10688 in Lecture Notes of Computer Sciences, pp. 189–201. Springer, 2017.

[27] HALL, M., F. EIBE, G. HOLMES, B. PFAHRINGER, P. REUTEMANN, and I. H. WITTEN: *The WEKA Data Mining Software: An Update*. *SIGKDD Explor. Newsl.*, 11(1), pp. 10–18, 2009.

[28] BREIMAN, L.: *Random Forests*. *Machine Learning*, 45(1), pp. 5–32, 2001.