# Speaker-Group Specific Acoustic Differences in Consecutive Stages of Spoken Interaction

*Ronald Böck*[1,2], *Olga Egorow*[1], *Andreas Wendemuth*[1,2]

[1]*Cognitive Systems Group, Otto von Guericke University Magdeburg,*
[2]*Center for Behavioral Brain Sciences*
*ronald.boeck@ovgu.de*
*www.cogsy.de*

**Abstract:** The communication of humans is influenced by various circumstances, especially in real-life situations. A particular aspect is the detection of affective states in an interaction and their changes. An automatic detection of such changes is challenging and highly related to significant features used for the detection. Therefore, we analysed the differences in spectral and prosodic features in an interaction, in particular, in two distinct communication stages. This investigation was done on a subset of 18 participants of the LAST MINUTE Corpus providing several different affect afflicted stages in a naturalistic Human-Computer Interaction. Across the 18 participants, 16 of the 52 analysed features – a subset of the *emobase* feature set – showed significant differences between the two considered stages. Most remarkably, seven of them are related to Mel-Frequency Cepstral coefficients. Furthermore, we found that there is a subset of five participants showing similar feature changes in the two stages. Four of them belong to the elder speaker-group, regardless of their sex. The presented findings confirm that features showing similar changes across and within speaker-groups as well as speaker-group dependent modelling are advantageous in the assessment of affective Human-Computer Interaction.

## 1 Introduction

Real-life human communication is affected by various circumstances, such as the current situation, the communication interlocutors, personal characteristics like sex and age, and so on. Furthermore, feelings, emotions, and dispositions (including for example being stressed or confused, etc.) are also possible influences. These aspects of Human-Human Interaction also occur in Human-Computer Interaction (HCI) (cf. [1]). On one hand, they make the interaction more challenging to process automatically, but on the other hand, they are an additional source of information. Using this additional input, interaction with technical systems can be made more human-like and better suited to the individual demands of the user. But to be able to use this information in HCI, first we need to assess and evaluate it.

In the study presented here we explore the influence of sex and age on the acoustic characteristics of speech in different stages of spoken close-to-real-life, naturalistic human-computer communication.

### 1.1 Research Questions

The aim of this study is to analyse the acoustic and prosodic differences during spoken interaction with a technical system. This is done in the context of a naïve, naturalistic communication, where the subject is (re-)acting in a non-scripted way. From literature (cf. Section 1.2) as well as from further analyses on the investigated data set (cf. Section 1.2 and 2, e.g. [2, 3, 4]) we know that in naturalistic interactions various situational stages can be distinguished by automatic classification. Currently, it is still a disputed issue which prosodic or spectral features provide the most impact in classifying those stage (cf. [5, 6, 7]). Below, we contribute to the discussion on feature sets and provide an insight on the discriminative power of particular features in a certain naturalistic HCI.

In particular, we are interested in the following three research questions:

Q1: Can spectral or prosodic features indicating differences in interaction stages across speakers be identified?

Q2: Can similar characteristics in the features' changes be identified related to particular speaker or speaker-groups?

Q3: Are the features' changes within particular speaker-groups related to similar features?

To answer these research questions, we investigated both, the features themselves as well as the shared characteristics within groups of speakers. For this, we currently selected speaker-groups as discussed in [3], namely age, sex, and corresponding combinations.

### 1.2 Related Work

Affect has been early recognised as an important aspect of HCI [8], with automatic recognition of human emotion – especially in natural situations – being one of the key challenges [9]. In the field of spoken communication, the focus also shifted from acted emotions, for example on the famous standard Berlin Database of Emotional Speech [10] to natural emotions or so called "in the wild" scenarios (cf. [11]). Although there have been many investigations concerning automatic emotion recognition (cf. [12, 13]), including investigations on the data we used for the investigations presented here (cf. [2, 3, 4]), there is still no consensus on which features are relevant for this task (cf. [7]), with a variety of feature sets being recommended, for example the *Geneva Minimalistic Acoustic Parameter Set* [5], different versions of the *emobase* feature set (cf. [14]), and so on. An additional problem is that acoustic emotion recognition using machine learning methods still does not provide satisfying results for "in the wild" scenarios. This poses the question, which features best describe the changes in the emotional content of spoken communication (cf. e.g. [7]). The effect of different emotional states on spectral features has been investigated for distinct emotions like fear, happiness, etc. for acted and pseudo-spontaneous emotions, where the actors were asked to act as naturally as possible (cf. [15, 16]). But, to the best of our knowledge, this question has not received enough attention for naturalistic scenarios such as in our case. Therefore, we investigated which spectral and prosodic features change and how exactly they change in different stages of close-to-real-life HCI – and whether there are differences with respect to the age and sex of the users.

## 2 Data Set: Last Minute Corpus

For this study, we used the LAST MINUTE Corpus (LMC, cf. [17]) – a collection of naturalistic HCI recordings during Wizard-of-Oz experiments. The interactions are divided in four distinct stages representing different situations a user can face (cf. [17, 18]). In the experiments, the users are first asked to pack a suitcase for a trip. After a while of packing, the users then discover that the suitcase has a weight constraint, and have to re-organise the suitcase. This event is one example for a so-called barrier (cf. [19]), that divides the interaction in several stages (cf. [18]). The barriers allow to align the users' utterances with a certain situation. In order to clarify the users' behaviour during the different stages, we give some examples. The first example shows an excerpt (cf. Table 1) from a stage consisting of "normal", untroubled interaction. The wizard (W) introduces a new item category and the user (U) starts choosing items.

**Table 1** – Excerpt of an untroubled interaction. The German version of the wizard's text was taken verbatim from [18].

| | German | | English |
|---|---|---|---|
| W: | *Sie können jetzt aus der Rubrik Hosen und Röcke auswählen* | W: | *You may now choose from the category trousers and skirts* |
| U: | *Zwei Jeans* | U: | *Two jeans trousers* |
| W: | *Zwei Jeans wurden hinzugefügt* | W: | *Two jeans trousers have been added* |

The second excerpt (cf. Table 2) comes from a more challenging stage after the so-called weight-limit barrier (cf. [19]) – the user now knows that the suitcase is full and tries to re-organise it.

**Table 2** – Excerpt of an interaction in a challenging stage. The German version of the wizard's text was taken verbatim from [18].

| | German | | English |
|---|---|---|---|
| W: | *Der Artikel Kosmetikset kann nicht hinzugefügt werden. Anderenfalls würde die maximale Gewichtsgrenze des Koffers überschritten werden* | W: | *The item cosmetic set cannot be added. Otherwise the weight limit of your suitcase will be exceeded* |
| U: | *Dann... einen ((leise)) ... ((klopft)) ... dann bitte einen Anorak raus* | U: | *Then... a... ((quietly)) ... ((knocks)) ... then please take out an anorak* |

In this study, we analysed the utterances of 18 participants (nearly equally distributed regarding sex and age, cf. Table 3) in the two previously introduced sub-scenarios: the more relaxed stage of packing and the more challenging stage of re-organising the suitcase. We selected the particular sub-group for this study since these participants did not only participate in the data collection of the LMC but also in two further experiments (cf. [20]). Therefore, we have the option to compare later our findings intrapersonally with analyses done on the other recordings.

**Table 3** – Distribution of sex and age over all considered users.

| | male | female | overall |
|---|---|---|---|
| young | 5 | 3 | 8 |
| elder | 6 | 4 | 10 |
| overall | 11 | 7 | 18 |

# 3 Experimental Setup

## 3.1 Feature Set

The utilised features for our investigations are based on the previously mentioned *emobase* feature set provided alongside the openSMILE feature extraction tool [14]. In general, *emobase* provides 988 prosodic and spectral features derived from functionals based on low-level descriptors (e.g. Mel-Frequency Cepstral coefficients (MFCC), intensity, loudness). Since we were interested in general trends, we decided to use mean values of features for our comparison. For this, we used the provided mean values contained in the *emobase* feature set, resulting in 52 features. Parts of these features are also known to be discriminative in emotion recognition from speech which provides a good relation between our analysis and a more general perspective on emotion recognition (cf. [5, 6, 21]) and emotion change detection from speech (cf. [22, 23]).

Finally, the mean values of features extracted on utterance-level were averaged over all utterances of the particular stage. In the end, we obtained a single value per feature per stage. The values of the first stage were then compared to the corresponding feature in the second stage (cf. Section 3.2).

## 3.2 Comparison Procedure

As stated in Section 3.1, we averaged the mean feature values over all utterances of a stage to allow a direct comparison of the inter-stage differences. In order to obtain interpretable results, we introduced as a comparative measure the difference $D$ of the corresponding features of the two stages. Given the particular mean feature $\overline{f}_i$ the measure is calculated as follows:

$$D = \overline{f}_i(s_1) - \overline{f}_i(s_2) \tag{1}$$

where $s_1$ and $s_2$ indicate the two affective stages.

We found that the difference allows a good interpretation of the results since changes in particular features can be directly assessed. Therefore, we decided to use this measure for further considerations. Nevertheless, also the ratio of the feature values can be considered as it provides a relational interpretion of the stages.

In addition to the $D$ values and the corresponding comparison, we computed correlation coefficients between the participants, namely the Pearson correlation coefficient $r$ (cf. [24]) and the Spearman correlation coefficient $\rho_s$ (cf. [25]).

Pearson correlation provides an assessment of the linear relation level between two characteristics (cf. [24]). In our investigations these are the LMC's participants. The $r$ value is between $-1$ and $+1$, corresponding to a negative correlation ($-1$), no correlation (0), and a positive (perfect) correlation ($+1$). Prerequisite for the Pearson coefficient is a (approximate) normal distribution of both measurements. In case of the participant's characteristics we cannot absolutely ensure this circumstance and thus calculated the Spearman coefficient, additionally.

Spearman's correlation coefficient $\rho_s$ (cf. [25]) can also be calculated even in cases where the data is not normally distributed. Generally, the Spearman coefficient compares the ranking of participant's measurements. This results in a numerical relation of rank changes.

# 4 Results

As described in Section 3.2, we calculated the differences of feature values. This was done for the 52 mean features presented in Section 3.1. For reasons of significance, we calculated the standard deviation for each feature across speakers as well as within speakers considering the corresponding difference values. Therefore, in this discussion we considered only those features and subjects, whose results lie outside the expected (and calculated) standard deviation.

Regarding research question Q1, we compared the changes in the various features across 18 participants (cf. Section 2). We found that 16 of 52 features showed significant differences between the two considered stages (cf. Table 4). In particular and most remarkably, 7 of these 16 features are MFCC-related. It is known from literature, that MFCCs are useful for classification of affected situations (cf. [21, 26, 27]). Our investigations point out the discriminative power of MFCCs not only in affect analysis but also in a broader sense of interaction stages. Further, it is of interest that higher-order Linear Spectral Pairs (LSP) – usually known from coding applications – were amongst the shared features. The advantage of spectral pairs in affect classification was analysed in, for instance, [28]. The interpretation of the observed higher-order LSPs is still a matter of discussion and is to be further elaborated in an extended study on the full LMC. Given these indications, the next step would be to conduct classification experiments of the two stages, comparing the full *emobase* feature set (cf. [14]) and the reduced set we employed here.

**Table 4** – Overview of 16 features showing a significant difference between two stages.

| | | | |
|---|---|---|---|
| Intensity | Loudness | MFCC 3...6 | MFCC 8...10 |
| LSP Frequency 2 | LSP Frequency 6&7 | Δ MFCC 5 | Δ LSP Frequency 1 |
| Δ LSP Frequency 5 | Δ F0 Envelope | | |

The second research question (cf. Q2 in Section 1.1) was related to characteristics in feature changes shared between particular speakers or speaker-groups. Based on the Spearman correlation coefficient $\rho_s$ we found that only two individual speakers are related to each other in terms of features ranking ($\rho_s = 0.5015$). For the remaining pairs no clear assessment was possible. In terms of Pearson correlation at least two individual speakers had an indication of correlation with eight and nine other speakers, respectively, mainly negatively correlated ($r < -0.5562$).
In contrast, observing speaker-groups we found that there was a subset of five subjects showing similar trends towards a decrease in the second stage. These differences also lay outside the expected standard deviation, significantly decreasing in the second stage. The most interesting aspect of this finding was that four of the five subjects belong to the elder speaker-group, regardless of their sex (two male and two female subjects). This is connected to the findings presented in [3], stating that speaker-group dependent modelling can improve the classification performance in affective HCI.

Regarding similarities in feature changes across speakers (cf. Q3 in Section 1.1), no clear results could be found. Therefore, we analysed the speaker-groups as proposed by Siegert et al. [3]: young/elder female and young/elder male, and found that the similarity in the feature changes were more homogeneous in certain speaker-groups.
All female speakers had a decreasing Δ F0 value comparing the two stages. Additionally, young female speakers had similar changes for 11 features, mainly related to LSPs and Δ LSPs, in particular in higher-orders. The LSPs showed a decrease in mean feature values while the Δ LSP values provided an increase. For elder female participants a mainly decreasing trend was also found for Δ LSPs features. Further, elder speakers had similar trends in feature changes for seven features in total.
For the male speaker-group also just one feature, namely MFCC 10, has a similar trend in changes across all participants. In general, for male subjects the indication is more divergent. In particular, elderly male participants seven features showed similar trends across the full speak-

ers' subset, mainly for MFCCs. In contrast, for young male subjects the similar feature changes spanned across all low-level descriptors, slightly focused on delta features. Therefore, no clear distinctive feature or feature group could be identified.

## 5 Conclusion and Outlook

In the current paper, we presented a comparison of spectral and prosodic features in two stages of a naturalistic interaction. The work was focussed on a subset of 18 participants of the LAST MINUTE Corpus (cf. [17]). This sub-group was choosen to allow analyses in two other experiments (cf. [20]). In our experiments, the analysed two stages are related to a normal untroubled interaction and a challenging one. In Section 3.2 we presented our comparison procedure based on the *emobase* feature set of openSMILE (cf. [14]). As explained in Section 4, we investigated the feature differences in two stage across features and speakers. We found relations between features and speaker-groups. The features showing similar and significant differences are mainly focused on MFCC and LSP values. For particular speakers, we also found significant trends towards a decrease of feature values in the second stage. We identified a particular group of speakers, namely four elder speakers, showing a similar trend in feature changes. For the speaker-groups discussed by Siegert et al. [3], the authors argue that speaker-group dependent modelling improves the classification performance in emotion recognition from speech across all sub-groups. From our findings we currently cannot state that this is related to similar trends in feature changes in particular speaker-groups.

In future work, we will extend the analyses to all speakers included in the LMC, expecting that the trends shown in this study can be assured. Additionally, we plan to relate the findings presented here to analyses of biological and textual features (cf. [20]) to investigate speaker characteristics multimodally. This is possible as the subset of 18 participants went through two other experiments examining biophysiological features and brain characteristics. Furthermore, classification experiments to investigate the opportunities provided by the identified features will be conducted.

## Acknowledgement

We acknowledge continued support by the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" (www.sfb-trr-62.de) funded by the German Research Foundation (DFG). Further, we thank the project "Mod3D" (grant number: 03ZZ0414) funded by 3Dsensation (www.3d-sensation.de) within the Zwanzig20 funding program by the German Federal Ministry of Education and Research (BMBF).

## References

[1] 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, New York, NY, USA, 1994.

[2] 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. Istanbul, Turkey, 2012.

[3] Siegert, I., D. Philippou-Hübner, K. Hartmann, R. Böck, and A. Wende-

MUTH: *Investigation of speaker group-dependent modelling for recognition of affective states from speech. Cognitive Computation*, 6(4), pp. 892–913, 2014.

[4] EGOROW, O. and A. WENDEMUTH: *Detection of challenging dialogue stages using acoustic signals and biosignals.* In *Proceedings of the WSCG 2016*, pp. 137–143. Plzen, Czech Republic, 2016.

[5] 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.

[6] SCHERER, K. R.: *Appraisal considered as a process of multi-level sequential checking.*, pp. 92–120. Oxford University Press, 2001.

[7] 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*, PP(99), pp. 1–14, 2015.

[8] PICARD, R. W.: *Affective computing for hci.* In *Proceedings of the International Conference on Human-Computer Interaction*, pp. 829–833. 1999.

[9] WARD, R. D. and P. H. MARSDEN: *Affective computing: problems, reactions and intentions. Interacting with Computers*, 16(4), pp. 707–713, 2004.

[10] 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. Lisbon, Portugal, 2005.

[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, 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, 2013.

[14] 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, New York, NY, USA, 2013.

[15] 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*. 2000.

[16] 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 (ICSLP 04)*. 2004.

[17] 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 Eight International Conference on Language Resources and Evaluation*, pp. 82–89. ELRA, 2012.

[18] 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.

[19] PANNING, A., I. SIEGERT, A. AL-HAMADI, A. WENDEMUTH, D. RÖSNER, J. FROMMER, G. KRELL, and B. MICHAELIS: *Multimodal affect recognition in spontaneous hci environment*. In *2012 IEEE International Conference on Signal Processing, Communication and Computing*, pp. 430–435. IEEE, Hong Kong, China, 2012.

[20] RÖSNER, D. F., D. HAZER-RAU, C. KOHRS, T. BAUER, S. GÜNTHER, H. HOFFMANN, L. ZHANG, and A. BRECHMANN: *Is there a biological basis for success in human companion interaction? - results from a transsituational study*. In *Human-Computer Interaction. Theory, Design, Development and Practice - 18th International Conference on Human-Computer Interaction*, LNCS, pp. 77–88. Springer, 2016.

[21] SCHULLER, B., B. VLASENKO, F. EYBEN, G. RIGOLL, and A. WENDEMUTH: *Acoustic Emotion Recognition: A Benchmark Comparison of Performances*. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 552–557. Merano, Italy, 2009.

[22] BÖCK, R. and I. SIEGERT: *Recognising emotional evolution from speech*. In *Proceedings of the International Workshop on Emotion Representations and Modelling for Companion Technologies*, pp. 13–18. ACM, New York, NY, USA, 2015.

[23] HUANG, Z., J. EPPS, and E. AMBIKAIRAJAH: *An investigation of emotion change detection from speech*. In *INTERSPEECH-2015*, pp. 1329–1333. ISCA, 2015.

[24] PEARSON, K.: *Note on regression and inheritance in the case of two parents*. *Proceedings of the Royal Society of London*, 58(347-352), pp. 240–242, 1895.

[25] SPEARMAN, C.: *The proof and measurement of association between two things*. *American Journal of Psychology*, 15, pp. 88–103, 1904.

[26] BÖCK, R., D. HÜBNER, and A. WENDEMUTH: *Determining optimal signal features and parameters for hmm-based emotion classification*. In *Proceedings of the 15th IEEE Mediterranean Electrotechnical Conference*, pp. 1586–1590. IEEE, Valletta, Malta, 2010.

[27] SCHULLER, B., A. BATLINER, S. STEIDL, and D. SEPPI: *Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge*. *Speech Communication*, 53(9-10), pp. 1062–1087, 2011.

[28] SHAHZADI, A., A. AHMADYFARD, K. YAGHMAIE, and A. HARIMI: *Recognition of emotion in speech using spectral patterns*. vol. 26, pp. 140–158. University of Malaya, 2013.