FILTERING-BASED ANALYSIS OF SPECTRAL AND TEMPORAL EFFECTS OF ROOM MODES ON LOW-LEVEL DESCRIPTORS OF EMOTIONALLY COLOURED SPEECH

Martin Gottschalk[1], *Juliane Höbel-Müller*[2], *Ingo Siegert*[3], *Jesko Verhey*[1], *Andreas Wendemuth*[2]

[1] *Department of Experimental Audiology,* [2] *Chair of Cognitive Systems,* [3] *Mobile Dialog Systems, Otto von Guericke University Magdeburg*
martin.gottschalk@med.ovgu.de

Abstract: Emotion recognition in far-field speech is challenging due to various acoustic factors. The present contribution especially considers dominant low-frequency room modes which are often found in small rooms and cause variations in the low-frequency acoustical response at various listening locations.
The impact of this spatial variation on low-level descriptors, used for feature sets in speech emotion recognition, has not been analysed in detail so far.
This shortfall will be addressed in this paper, by utilising the well-known benchmark dataset EMO-DB providing emotionally coloured speech of high quality. The measured room response of a speaker cabin is compared with artificial approximations of its frequency response in the low frequency range. Two techniques were applied to obtain the approximations: The first technique uses multiple resonant filters in the low frequency region, whose parameters are determined by a least-squares fit. The second technique used a modified version of the cabin's amplitude spectrum, that was set to unity for higher frequencies and transformed to minimum phase and to time domain.
To be able to identify the impact of room modes on the low-level descriptors, correlation coefficients between the "clean" and modified EMO-DB utterances are calculated and compared to each other. Furthermore, a speech emotion recognition system is used to identify the impact on the recognition performance.

1 Introduction

Voice-based human-machine interaction (HMI) "in the wild" is exposed to varying environment conditions. It has been analysed in terms of superposed noise [1, 2], robust feature sets [3, 4], feature pooling [5] or feature degradation for different room acoustics [6] and their impact on emotion recognition performance [7]. Furthermore, the impact of room acoustic characteristics on specific feature types and the performance of speaker state classification has been analysed [8, 9, 10]. It could be shown that emotion recognition in far-field speech shows performance drops due to several environmental factors, including background noise, echo, reverberation, delay and other.

One factor that has been neglected so far are dominant low-frequency room modes which are often found in small rooms and cause variations in the low-frequency acoustical response at various listening locations. This spatial variation impacts the speech signal. Also the low-level descriptors (LLDs) used in various feature sets for speech emotion recognition are affected. Therefore, speech emotion recognition may be challenging, for instance, in Ambient Assisted Living environments, as user's far-field voice is necessary due to user acceptance considerations.

The present work analyses the effects of low-frequency room modes as an environmental factor on the recognition of emotionally coloured speech. Therefore, a room impulse response (RIR) of an acoustically damped speaker cabin representing a small room where all other factors except the low-frequency room modes are suppressed, is measured. Afterwards, different approximations to the measured RIR are synthesized in order to reduce the temporal effects while maintaining the frequency response. Both, the originally measured RIR and its approximated variants are convoluted with acoustic signals from a benchmark dataset of emotional speech, to obtain speech degraded by the acoustic properties of our small room.

Afterwards, the extracted LLDs for the different variants are compared and the influence of the RIR and its approximations are discussed. To furthermore draw a conclusion regarding the influence on an emotion recognition system, identical classification experiments are conducted for both measured and approximated RIRs.

2 Room Modes

The acoustic resonances of standing waves between parallel walls are called room modes. In general, they cause non-uniform sound fields with large differences in sound pressure depending on the listening position relative to the nodes and antinodes of a room mode. This results in a nonuniform frequency response. In typical living rooms, isolated room modes appear mainly below 200 Hz. In smaller rooms (e. g. bathroom, phone booth), isolated room modes can occur in a higher frequency range, including fundamentals and formants (F1) of speech.

The frequency of a longitudinal standing wave is given by

$$f = \frac{c}{2}\frac{k}{L}$$

where k is the order of the room mode, c is the acoustic velocity, and L is the respective room dimension (e.g length). Furthermore, two-dimensional tangential modes and three-dimensional oblique modes exist, but are typically less relevant. With room length L, width W, height H and mode orders k, m, n, their frequencies are given by

$$f = \frac{c}{2}\sqrt{\left(\frac{k}{L}\right)^2 + \left(\frac{m}{W}\right)^2 + \left(\frac{n}{H}\right)^2}.$$

In this study, a small recording booth was used. Its most prominent room modes were the longitudinal modes of the identical length and width dimension at 76 Hz (first order) and 152 Hz (second order), the longitudinal mode of the height dimension at 138 Hz (second order) and 202 Hz (third order).

3 Experimental Set-up

3.1 Emotional Speech Corpus

To enable a valid ground truth and guarantee high quality recordings we utilized the Berlin Database of emotional Speech (EMO-DB) [11]. This dataset consists of German utterances with neutral semantic content, uttered by five female and five male professional actors in the seven basic emotions anger, boredom, disgust, fear, joy, neutral, and sadness. The female speakers were on average 30.6 $\pm$ 5.6 and the male speakers were on average 28.8 $\pm$ 3.1 years old. Overall, the database contains 494 utterances spanning between 2 and 5 seconds.

The high quality samples were originally recorded in an anechoic chamber using a Sennheiser MKH 40-P48 microphone with a sampling frequency of 48 kHz, later downsampled to 16 kHz.

In a perception test, conducted by the corpus creators, all samples below 60% naturalness and 80% emotion recognizability were discarded, resulting in 494 phrases. Unfortunately, due to the removal of several recordings, the gained distribution of emotional samples is unbalanced.

3.2 Measuring the RIR

In order to measure the RIR, we used a hardware setup, which is characterized by a highly linear frequency response. In particular, we used a Behringer ECM8000 ultra-linear condenser microphone with omnidirectional pattern, a Yamaha 01V96i audio interface and a Neumann KH120A loudspeaker. We applied CARMA Version 4.0, a room acoustics analysis program, to determine the RIR and the resulting amplitude response, based on a logarithmic sinusoidal signal in 44.1 kHz as the measuring stimulus. The sinusoidal signal was played and re-recorded in a small room with the dimensions 2.22 m $\times$ 2.22 m $\times$ 2.44 m (length $\times$ width $\times$ height).

By Matlab-based convolving each EMO-DB utterance with a RIR of our small room, we obtained emotionally expressive speech, which is degraded by the acoustics of the room. This degradation can be considered as a combination of primarily temporal effects (reflections, reverberation) and primarily spectral effects. In order to separate spectral and temporal impacts, the temporal impacts were mitigated. This allows to compare the results with the original room, so that the effects of spectral and temporal component can be compared.

3.3 Mitigating Room Modes' Temporal Component's Effect on Speech

The RIR was separated into a low-frequency and a high-frequency part, assuming that room modes are not relevant for frequencies higher than 250 Hz. A crossover filter with 24 dB/octave was used to achieve that. The high-frequency part contained mainly comb filtering in the frequency space, which corresponds to reflections in the time space. The high-frequency part was replaced by a unity response to remove those reflective characteristics. Merely high-frequency roll-off was applied to emulate the original high-frequency roll-off of the RIR due to the recording equipment and wall damping. The low-frequency part was treated in three different ways, corresponding to the three investigated conditions.

a) Multiple Infinite Impulse Response (IIR) Filters were used to emulate the spectral properties of the original RIR. For this purpose, a set of filters was chosen that is sufficient to replicate the rough structure of the amplitude spectrum in the low-frequency part. The parameters of these filters were fitted to the amplitude spectrum using a nonlinear least-squares algorithm, after a 1/3-octave wide spectral smoothing. In the case of the studied RIR, two high-pass filters and five peaking EQ filters, also known as bell filters, were used, each one of second order. The combination of these filters was converted to time space. This way, we obtained an artificial RIR that approximates the spectral changes by the room modes without reflections or reverberation. This condition will be referred to as "Filters" in the following.

b) The low-frequency part of the RIR was converted to a minimum phase (MP) response by an operation in the complex cepstrum space. The right side of the cepstrum corresponds to maximum-phase zeros. Flipping the right side and adding it onto the left side leads to a minimum-phase cepstrum. This was inversely transformed back to the time space. The amplitude response was not changed by that operation. This condition will be referred to as "Low-freq MP" in the following.

c) The low-frequency part of the RIR was not changed. This means that all temporal aspects of the low-frequency response were conserved. This condition will be referred to as

"Low-freq" in the following.

3.4 Extracting Low Level Descriptors (LLDs)

We extracted 26 low-level descriptors (LLDs) from the clean EMO-DB utterances as well as from the utterances convoluted with the original RIR and with the artificial ones. By applying the openSMILE toolkit [12], the LLDs were extracted on a 25 ms frame-level regarding the benchmark emobase configuration. In particular, these LLDs belong to loudness-, cepstral-, LPC-, waveform- and pitch-related feature groups.

3.5 Extracting Spearman's Correlation Coefficient for Low Level Descriptors (LLDs)

We intended to compare a clean utterance with a time-aligned synthesised one (cf. 3.3). As both utterances originated by the same speaker, we would assume a linear association between them and consequently between their LLDs. Visual examination of randomly chosen scatter plots seemed to confirm this assumption. The Pearson product moment correlation r_p coefficient is a measure in order to estimate the degree of *linear* association between two variables [13]. Assuming measurements on two LLDs X and Y for n samples, the paired LLD values can be written as (x_1, y_1), (x_2, y_2), ... , (x_n, y_n), where x_i, y_i should be normally distributed. Then, the sample Pearson product moment correlation coefficient, r_p, is defined as

$$r_p(X,Y) = \frac{\mathbf{Cov}(X,Y)}{\sqrt{\mathbf{Var}(X)\mathbf{Var}(Y)}}. \quad (1)$$

As r_p relies on normally distributed LLD values, we tested our data for normal distribution. One-sample Kolmogorov-Smirnov tests [14] rejected the null hypothesis that a LLD comes from a standard normal distribution at a corrected 5 % significance level. In contrast, the Spearman rank correlation coefficient r_s, does not require the assumption of normality, which motivated us to apply it. r_s was obtained by ranking the values of two LLDs and calculating the Pearson correlation coefficient r_p and the population value by p_s, on the resulting ranks. Equation 1 may be used to calculate r_s if (x_i, y_i) are replaced by their ranks (r_i, s_i) [13].

3.6 Emotion Recognition Experiments

To perform the emotion recognition experiments, state-of-the-art automatic recognition experiments comparable to [15] are conducted. In contrast to them, it is opted for a Leave-One-Speaker-Out (LOSO) validation scheme to better represent realistic applications.
For feature extraction, the same feature set as for the statistical analyses has been used, with the only difference that the functionals are used, resulting in 988 features characterizing the supersegmental distribution per utterance. Afterwards, standardization as normalisation technique is used to eliminate differences between the data samples [16]. As recognition system, a SVM with linear kernel and a cost factor of 1 was utilized with WEKA [17]. As performance measure, the F-measure (FM) were calculated as the average over the single speaker's performance measure.

4 Results

4.1 Analysis of Room Impulse Responses

The RIR of the speaker cabin and the artificial RIRs as described in section 3.3 are shown in Fig. 1. In Fig. 1 a), all RIRs show a main peak corresponding to the direct sound (partly not visible behind the green line). The "Original" RIR also shows several smaller peaks corresponding

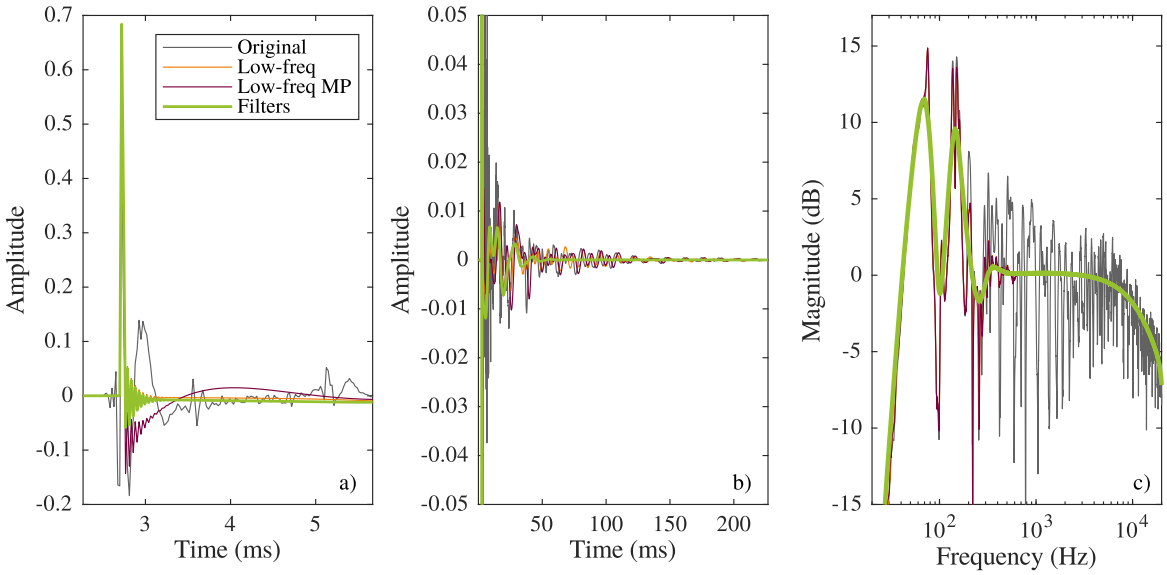

Figure 1 – The original RIR (in grey color) in comparison to the three artificial RIRs (Low-freq in yellow color, Low-freq Minimum Phase in dark red color and the fitted Filters in green color). Fig. 1 a) and b) show time domain representations and Fig. 1 c) shows a frequency domain representation of the same RIRs.

to reflections for example from the side wall approximately 2.4 ms after the main peak. The artificial RIRs do not show these reflections, as discussed in section 3.3. A longer section of the RIRs is shown in Fig. 1 b). The plot shows, that the decay of the "Original", "Low-freq" and "Low-freq MP" RIRs is similar, whereas the decay of the "Filters" RIR is much faster than the others. Figure 1 c) shows, how the impact of all three artificial RIRs ("Low-freq", "Low-freq MP" and "Filters") is limited to the lower frequency range roughly below 250 Hz. Also, the high-frequency roll-off due to the recording equipment and wall damping around 10 kHz is visible.

4.2 Analysis of Correlation Coefficients

In the first part of our statistics-related analysis, we answer the question how many correlation coefficients significantly differ from zero, with regard to the four experimental conditions "Original", "Low-freq", "Low-freq MP" and "Filters". So, in each experimental condition, we analyse 26 LLDs $\times$ 494 utterances resp. 12844 Spearman's correlation coefficients to quantify the strength of association between "clean" LLD series from clean speech and the corresponding LLD series from the four above mentioned experimental conditions. Note that the "Original" condition refers to the LLD series corresponding to EMO-DB, which was convolved by the speaker cabin's RIR, and *not* refers to clean speech. By examining the single Spearman's correlation coefficients in the four experimental conditions, we observed a slightly increasing number of LLD series, which strongly correlate with the "clean" LLD series. In particular, 96 % in the "Original" condition, 97 % in the "Low-freq" condition, 97 % in the "Low-freq MP" and 98 % in the "Filters" condition significantly differ from $r_s = 0$ at the Bonferroni-corrected 5 % significance level.

In the second part of our analysis, we answer the question of how strong is the correlation between "clean" LLD series and LLD series originating from the four conditions. To answer this question, we statistically analyse the distribution of coefficients, which significantly differ from zero (cf. previous question), by using box plots. These are generated by using Matlab. In Fig. 2, box plots [18] show the distribution of the Spearman's correlation coefficients across the four experimental conditions "Original", "Low-freq", "Low-freq MP" and "Filters".

An increasing median r_s across the conditions can be observed. We obtained an increasing

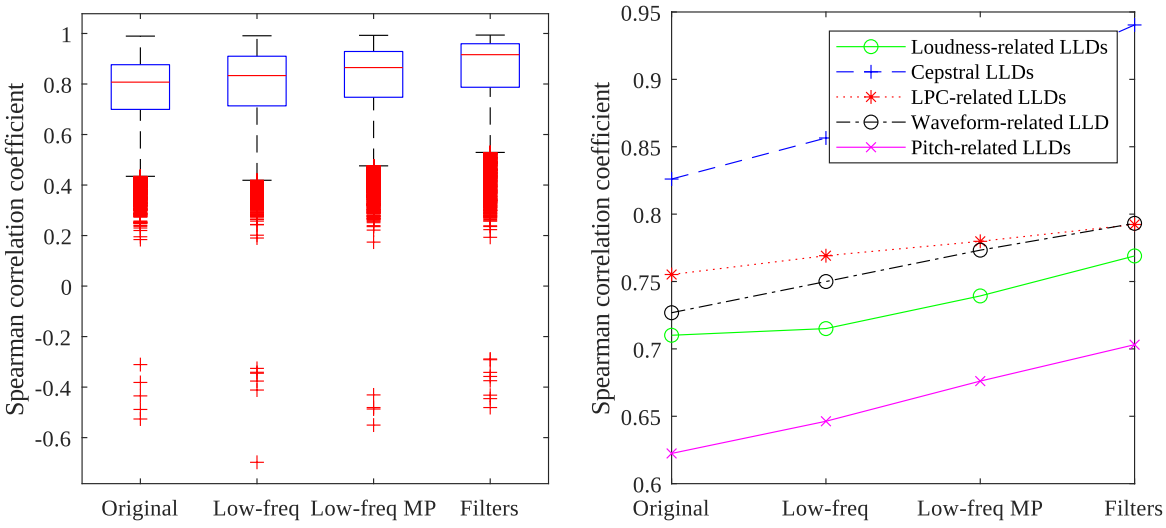

Figure 2 – **(a)** Boxplots for depicting sets of Spearman's correlation coefficients. Each coefficients' set corresponds to one experimental condition described in Sec. 3.3. **(b)** A glimpse of the association strength between loudness-, cepstral-, LPC-, waveform-, and pitch-related LLD series.

median of 0.81 in the "Original", 0.83 in the "Low-freq", 0.86 in the "Low-freq MP" and 0.92 in the "Filters" condition. Regarding the upper and lower quartiles, the correlation coefficients deviate similarly. Every condition contains outliers representing negative correlation coefficients in the range from $r_s = -0.3$ to $r_s = -0.7$. In particular, 3.2 %, 2.6 %, 4.1 % and 5.2 % of the correlation coefficients across the different conditions are outliers.

In the third part of our analysis, we give a glimpse with respect to the strength of association between the LLD series regarding the emobase feature group's loudness, cepstral, LPC, waveform, and pitch. In order to do that, we averaged the corresponding correlation coefficients per feature group and summarised the results in Fig. 2 b). In Fig. 2 b) one can see the trend of increasing LLD correlation coefficients per feature group, which we have already indicated before in Fig. 2 a). Additionally, the cepstral-related correlation coefficients clearly differ from the LPC-, the waveform-, and, last, the pitch-related ones.

Table 1 – Emotion recognition performances (F_1 score) for different experimental conditions and significance measures. The identifiers correspond to the conditions presented in Sec. 3.3 or the ones presented in Fig. 1.

Identifier	F_1 (std) [%]	Significance level
Clean EMO-DB (baseline)	78.18 (0.631)	–
Original (degraded by RIR)	73.63 (0.104)	$p < 0.001$
Low-freq	73.60 (0.706)	$p < 0.001$
Low-freq MP	75.69 (0.789)	$p < 0.001$
Filters	74.97 (0.654)	$p < 0.001$

In the last part of our analysis, we present emotion recognition performances (F_1) based on standardised emobase feature values, Support Vector Machines (SVMs) and a LOSO evaluation design. Our baseline is represented by the emotion recognition performance in clean speech. In Tab. 1, one can see the F_1 values for each experimental condition. As expected, the best performance is obtained for the baseline, which is followed by the "Low-freq MP" and "Filters" condition. These results in turn are followed by the "Original" and, last, by the "Low-freq" condition. By using a one-sided ANOVA, we show that the recognition results differ significantly.

5 Discussion and Conclusion

The room modes analysed in this work only have a small impact on the emotion recognition performance measure compared to the baseline. Increasing emotion recognition performances come slightly along with the mitigation of room modes' temporal effect in speech. In the "Filters" condition, the spectral changes by the room modes without *reflections or reverberation* were approximated, whereas all temporal aspects of the low-frequency room response were conserved in the "Low-freq" condition. The F_1 values for the "Filters" and "Low-freq" condition suggest a larger temporal-related impact on speech, however the results only slightly differ.

The increasing F_1 values are slightly accompanied by the correlation-based results, which are presented in Fig. 2. The slight descent in recognition performance for the last "Filters" condition is not in accordance with the fact that the aggregated correlation coefficients are highest compared to the other conditions. Another factor, the increasing number of negative correlation coefficients (outliers) across the conditions mentioned, comes along with the recognition results. In particular, 3.2 %, 2.6 %, 4.1 % and 5.2 % of the correlation coefficients across the different conditions are outliers. Obviously, the first two and the last two percentages are located close together and seem to form two clusters. Two clusters can also be observed in the recognition results presented in Tab. 1. Both in the outlier and recognition results clustering, we can observe the same experimental conditions pairs. Due to this similarity in condition membership, one can conclude that the recognition results are accompanied by the positive and negative correlated LLD series in combination.

In total, the effect of the room acoustics of the speaker cabin on the emotion recognition performance were small across all conditions (no larger than 5 % difference). Minimum-phase conditions ("Filters", "Low-freq MP") showed a slightly better performance than non-minimum phase conditions. This could imply, that emotion recognition is more robust against spectral than temporal effects of room acoustics. However, the obtained differences are too small to draw firm conclusions. Regarding far-field applications, an automatic analysis of a sinus sweep could give a glimpse of feature group-related distortions and motivate context-based feature selection.

References

[1] SCHULLER, B., D. ARSIC, F. WALLHOFF, and G. RIGOLL: *Emotion recognition in the noise applying large acoustic feature sets*. In *Proc. Speech Prosody 2006, Dresden*. 2006.

[2] TAWARI, A. and M. M. TRIVEDI: *Speech Emotion Analysis in Noisy Real-World Environment*. In *2010 20th International Conference on Pattern Recognition*, pp. 4605–4608. 2010.

[3] KIM, E. H., K. H. HYUN, and Y. K. KWAK: *Robust emotion recognition feature, frequency range of meaningful signal*. In *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.*, pp. 667–671. 2005.

[4] LEE, K.-K., Y.-H. CHO, and K.-S. PARK: *Robust Feature Extraction for Mobile-Based Speech Emotion Recognition System*. In D.-S. HUANG, K. LI, and G. W. IRWIN (eds.), *Intelligent Computing in Signal Processing and Pattern Recognition: International Conference on Intelligent Computing, ICIC 2006 Kunming, China, August 16–19, 2006*, pp. 470–477. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[5] AVILA, A. R., Z. A. MOMIN, J. F. SANTOS, D. O'SHAUGHNESSY, and T. H. FALK:

Feature Pooling of Modulation Spectrum Features for Improved Speech Emotion Recognition in the wild. *IEEE Transactions on Affective Computing*, pp. 1–1, 2018.

[6] HÖBEL-MÜLLER, J., I. SIEGERT, R. HEINEMANN, A. F. REQUARDT, M. TORNOW, and A. WENDEMUTH: *Analysis of the influence of different room acoustics on acoustic emotion features*. In *Elektronische Sprachsignalverarbeitung 2019. Tagungsband der 30. Konferenz*, pp. 156–163. Dresden, Germany, 2019.

[7] HÖBEL-MÜLLER, J., I. SIEGERT, R. HEINEMANN, A. F. REQUARDT, M. TORNOW, and A. WENDEMUTH: *Analysis of the influence of different room acoustics on acoustic emotion features and emotion recognition performance*. In *Tagungsband - DAGA 2019*, pp. 886–889. Rostock, Germany, 2019.

[8] AHMED, M. Y., Z. CHEN, E. FASS, and J. A. STANKOVIC: *Real Time Distant Speech Emotion Recognition in Indoor Environments*. In *MobiQuitous*. 2017.

[9] SCHULLER, B.: *Affective speaker state analysis in the presence of reverberation*. *International Journal of Speech Technology*, 14(2), pp. 77–87, 2011.

[10] EYBEN, F., F. WENINGER, and B. SCHULLER: *Affect recognition in real-life acoustic conditions-a new perspective on feature selection*. In *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*. 2013.

[11] BURKHARDT, F., A. PAESCHKE, M. ROLFES, W. SENDLMEIER, and B. WEISS: *A Database of German Emotional Speech*. In *Proc. of the Interspeech-2005*, pp. 1517–1520. Lissabon, Portugal, 2005.

[12] EYBEN, F., M. WÖLLMER, and B. SCHULLER: *openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor*. In *Proc. of the 18th ACM International Conference on Multimedia*, MM '10, pp. 1459–1462. ACM, New York, NY, USA, 2010.

[13] SPRENT, P. and N. C. SMEETON: *Applied Nonparametric Statistical Methods*. Chapman and Hall/CRC, 2001.

[14] MASSEY, F. J.: *The Kolmogorov-Smirnov Test for Goodness of Fit*. *Journal of the American Statistical Association*, 46(253), pp. 68–78, 1951.

[15] LEFTER, J., H. NEFS, C. JONKER, and L. ROTHKRANTZ: *Cross-corpus analysis for acoustic recognition of negative interactions*. In *Proc. of the 6th ACII*, pp. 132–138. Xian, China, 2015.

[16] BÖCK, R., O. EGOROW, I. SIEGERT, and A. WENDEMUTH: *Comparative Study on Normalisation in Emotion Recognition from Speech*. In P. HORAIN, C. ACHARD, and M. MALLEM (eds.), *Proc of the 9th IHCI 2017*, pp. 189–201. Springer International Publishing, Cham, 2017.

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

[18] TUKEY, J. W.: *Box-and-whisker plots*. *Exploratory data analysis*, pp. 39–43, 1977.