Investigation of the influence of standing waves on distant speech emotion recognition

Juliane Höbel-Müller[1,2], Ingo Siegert[3], Martin Gottschalk[4], Ralph Heinemann[1], Andreas Wendemuth[1]
[1] *Cognitive Systems Group, Otto von Guericke University Magdeburg,*
[2] *Data and Knowledge Engineering Group, Otto von Guericke University Magdeburg,*
[3] *Mobile Dialog Systems, Otto von Guericke University Magdeburg,*
[4] *Department of Experimental Audiology, Otto von Guericke University Magdeburg*
corresponding author: juliane.hoebel@ovgu.de

Introduction

Emotion recognition in far-field speech is challenging due to various acoustic factors. So far, conducted analyses have revealed that emotion recognition performances in far-field conditions drop due to several environmental factors, including background noise, echo, reverberation and other [1, 2, 3, 4]. Room modes, also known as acoustical resonances, represent a previously neglected factor. They lead to nonuniform sound pressure levels depending on position and frequency. Small rooms (e.g. bathrooms, cars) have their fundamental resonances in the range of speech. These impact both the speech signal and the low-level descriptors (LLDs) used in various feature sets for speech emotion recognition [5].

This work analyses the effects of low-frequency room modes as an environmental factor on the recognition of emotionally coloured speech in a real recording environment. We measured a room impulse response (RIR) of an acoustically damped speaker cabin and a RIR of an absorber hall. Due to the absorber panels in the cabin, all environmental factors except room modes in the range of speech were suppressed, whereas the hall represented free-field conditions and room modes outside of the range of speech. Additionally, artificial alterations based on the measured RIRs were computed. This was aimed to shorten the decay time while maintaining the low-frequency magnitude response of the originals.

The analyses were based on the benchmark dataset Berlin Database of emotional Speech (EMO-DB). Degradation due to the rooms was applied to the complete data set. The recorded data was convolved with both the original and altered RIRs of the cabin and the hall. The LLDs were extracted for the different variants and compared with the original EMO-DB. The influence of the measured and artificial RIRs was discussed. To furthermore draw a conclusion regarding the influence on an emotion recognition system, we have conducted identical cross-variant classification experiments for both measured and altered RIRs. The room modes' impact can be attributed to up to 6 % loss in F_1-measure. The altered RIRs lead to better feature and recognition performances.

Experimental Setup

In the following, the experimental setup is described. This includes the introduction of the used benchmark dataset of emotional coloured speech, the utilized measuring equipment, and the investigated rooms.

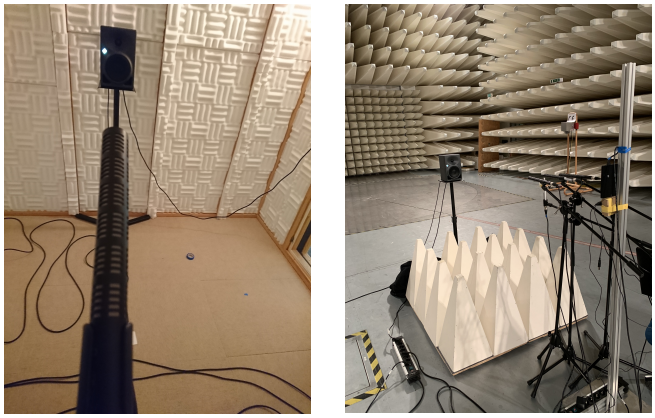

Figure 1: Photos of the used recording setup: Speaker cabin (left) Absorber hall (right).

Emotional Speech Data: The Berlin Database of emotional Speech (EMO-DB) [6] was utilized to guarantee high-quality recordings and enable a valid ground truth. EMO-DB consists of German utterances with neutral semantic content, uttered by five female and five male professional actors in seven basic emotions (anger, boredom, disgust, fear, joy, neutral, and sadness). The samples were originally recorded in an anechoic chamber using a Sennheiser MKH 40-P48 microphone at 48 kHz sampling frequency, and later down-sampled to 16 kHz. As shown in a figure in [6], the actors stood during the recordings. The microphone distance was about 30cm. In a perception test, conducted by the corpus creators, all samples below 60% naturalness and 80% emotion recognisability were discarded, resulting in 494 phrases. Unfortunately, due to the removal of several recordings, the gained distribution of emotional samples was unbalanced.

Measuring Equipment: In order to measure the RIR, a hardware setup characterized by a highly linear frequency response was used: a Behringer ECM8000 ultra-linear condenser microphone with an omnidirectional pattern, a Yamaha 01V96i audio interface and a Neumann KH120A loudspeaker. The RIR and the resulting amplitude response was then determined through CARMA Version 4.0, a room acoustics analysis program, utilizing an exponential sinusoidal sweep signal in 44.1 kHz as the measuring stimulus.

Investigated Rooms: To analyse the influence of standing waves, two rooms representing the extreme conditions regarding standing waves were selected. First, an acous-

tically damped speaker cabin representing a small room where all other factors except the low-frequency room modes are suppressed, and second, an absorber hall representing free-field conditions and thus not having isolated room modes in the frequency range of speech, see Fig. 1. The dimensions of the speaker cabin were 2.22 m × 2.22 m × 2.44 m (length × width × height), and the ones of the absorber hall were 21 m × 13 m × 9 m (length × width × height). Based on the dimensions of the room and the placement of the microphone and speaker, the lowest longitudinal room mode frequencies in the speaker cabin were expected to be 76.6 Hz (fundamental mode of the width and length dimension) and 139.3 Hz (second-order mode of the height dimension). In the absorber hall, the fundamental longitudinal modes' frequencies were expected to be 8.1 Hz (length), 13.1 Hz (width) and 18.9 Hz (height), respectively.

Methods

We used the following two methods in order to obtain speech data degraded by the rooms:

i) Each EMO-DB utterance was convolved with the RIRs of the speaker cabin ("Cabin") and the absorber hall ("Hall"), measured with the same measuring equipment, see Section "Experimental Setup". The convolution was based on Matlab.
ii) To attain minimum phase (MP) conditions, EMO-DB was convolved with artificial alterations of the measured RIRs of the speaker cabin ("Cabin MP") and the absorber hall ("Hall MP"). The alterations featured shorter decay time, identical low-frequency spectrum and minimum phase characteristics (see next section).

The resulting variants of EMO-DB are given in Table 1.

Calculation of artificial Minimum Phase RIR: In order to disentangle the influence of temporal and spectral effects, an artificial RIR was created based on the measured RIR of the speaker cabin and absorber hall, respectively. To maintain the low-frequency magnitude response of the original, but with a faster temporal decay, the following changes were made: i) the higher-frequency part of the spectrum was set to unity in the frequency space to suppress comb filtering, and ii) the RIR was converted to minimum phase by an operation in cepstral space.

Low Level Descriptor Extraction: For each variant of EMO-DB, 26 LLDs were extracted, applying the openSMILE toolkit [7]. We utilised the emobase configuration, defining 25 ms frame-level for the windowing of the speech utterance. This configuration contains LLDs belonging to loudness-, cepstral-, LPC-, waveform- and pitch-related feature groups.

Correlation of LLDs: The utterances of the clean EMO-DB were compared with the identical time-aligned utterances of the artificial ones. As both utterances originated from the same speaker, we would assume a linear association between them and consequently between their LLDs. The extracted LLDs are not normally distributed.

Table 1: Overview of generated EMO-DB variants.

Identifier	Description
EMO-DB	Original EMO-DB
Speaker Cabin's RIR convolved with EMO-DB:	
Cabin	Measured RIR of speaker cabin
Cabin MP	Artificial alteration (smoothed and minimum phase)
Absorber hall's RIR convolved with EMO-DB:	
Hall	Measured RIR of absorber hall
Hall MP	Artificial alteration (smoothed and minimum phase)

So, we used the Spearman rank correlation coefficient r_s, as it does not require the assumption of normality.

We used the Spearman rank correlation coefficient r_s, as it does not require the assumption of normality. In Matlab, r_s was obtained by ranking the values of two LLDs, and calculating the Pearson correlation coefficient and the population value, on the resulting ranks [8]. For our analysis, we considered only correlation coefficients, which differed significantly from $r_s = 0$ at a Bonferroni-corrected 5 % significance level.

Emotion Recognition Experiments: Finally, state-of-the-art automatic recognition experiments comparable to [9] were conducted. To check how well our learned models generalize to testing data, we opted for a Leave-One-Speaker-Out validation scheme. Furthermore, the training was conducted on the original EMO-DB, while for the test, we used different variants (cross-variant experiments), given in Table 1.

The feature extraction relied on the same feature set as the correlation experiments with the only difference to using the functionals instead of the LLDs, resulting in 988 features characterizing the super-segmental distribution per utterance. Afterwards, normalisation (standardization) was used to eliminate differences between the data samples [10]. As a recognition system, SVMs with linear kernel and a cost factor of 1 were utilized with WEKA [11]. As a performance measure, the F-measure (F_1) was calculated as the average over the single speakers.

Results

In the following, the results for the different analyses are presented. First, measured and artificial RIRs are presented. Afterwards, the correlation between the original EMO-DB and the generated variants is calculated, based on the extracted LLDs. Finally, the impact on emotion recognition is shown.

Analysis of Room Impulse Responses: Figure 2 shows measured impulse responses of the speaker cabin (measured data in grey colour). The original RIR's spectral representation of the speaker cabin in Figure 2b) features substantial variations of magnitude in the low-frequency region due to the room modes in that frequency region, whereas the RIR in Figure 2d), measured in the

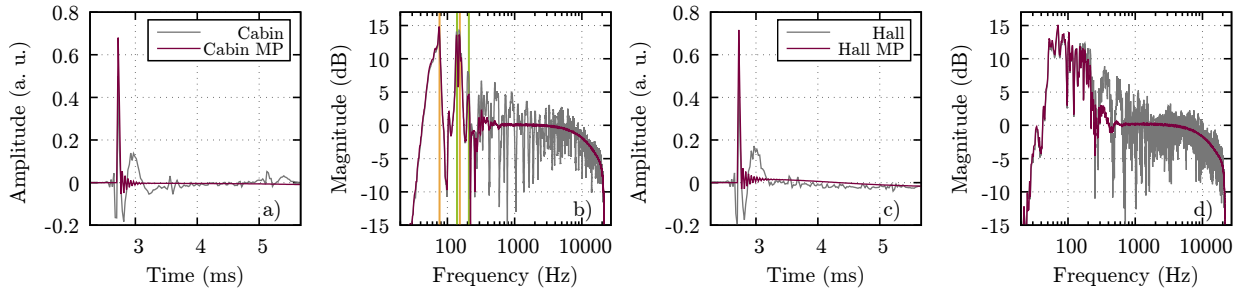

Figure 2: Original (grey) and altered (purple) versions of the speaker cabin (subfigures a), b)) and absorber hall (subfigures c), d)) RIR in a temporal and spectral representation. The yellow lines in subfigure b) indicate the expected frequency of the fundamental and second order room mode of the length and width dimension. The green lines indicate the expected frequency of the second and third-order room modes of the height dimension.

absorber hall, shows a much more uniform response in that frequency region. In the temporal representation, it is visible that the early parts of the RIRs of the speaker cabin in Figure 2a) and the absorber hall Figure 2c) are very similar because this part is determined by the transfer characteristics of the recording equipment. However, the first echo of the speaker cabin is visible around the 5 ms mark in the plot, whereas the first echo of the RIR absorber hall is much later than that, which is not shown here. Note that the impulse responses are considerably longer than the part shown in the plot and that the "MP" conditions (purple lines) are not zero outside of the shown range, as it might seem from the plots. The higher-frequency parts of the "MP" conditions' (purple lines) spectra are smoothed, so that the comb filtering in that range is eliminated.

Analysis of RIR's Influence on Emotion Features:

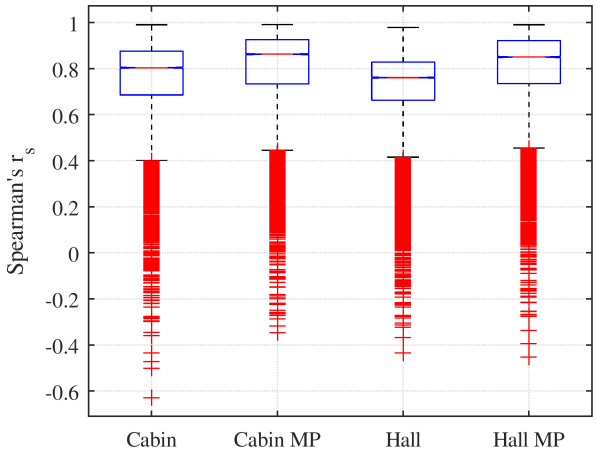

Figure 3: Notched boxplots show the distribution of Spearman's r_s for each experimental condition.

97 % of the correlation coefficients in the "Cabin", 97.5 % in the "Cabin MP", 96.5 % in the "Hall" and 97.2 % in the "Hall MP" condition differ significantly from $r_s = 0$ at the Bonferroni-corrected 5 % significance level, which means that these are significantly correlated with the clean LLDs, positively or negatively. Applying the boxplots in Figure 3, we can compare the range and distribution of the Spearman's correlation coefficients of all experimental conditions. The boxes indicate the median and

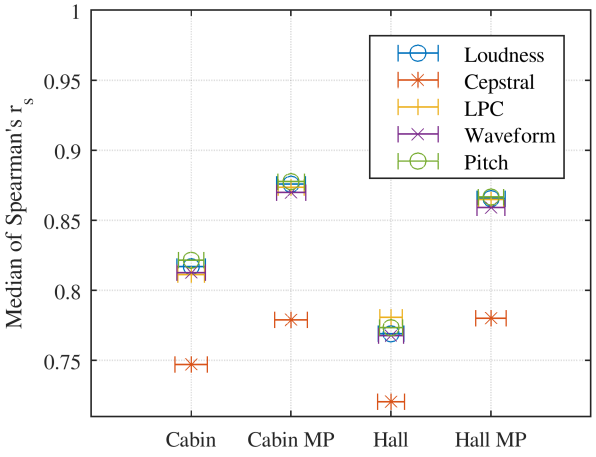

Figure 4: Median and standard deviation of Spearman's r_s per feature group and per experimental condition.

interquartile range. The length of the whiskers indicates the highest/ lowest value inside 1.5 times the interquartile range. The '+' indicates a value lying outside that range (outlier).

We observe that the coefficients vary similarly across all experimental conditions. Moreover, we observe a right-skewed distribution across all conditions, which means it has a few relatively low correlation coefficients. As the notches in the boxplots do not overlap, we conclude that the medians differ significantly on a 5 % level. The "Cabin"'s and "Cabin MP"'s medians are slightly higher than the "Hall" and "Hall MP" ones. Across all conditions, the "Cabin" condition shows a few more massive outliers.

Using the median and standard deviation of the correlation coefficients per feature group, we can compare feature-specific statistics in Figure 4. All medians deviate similarly in the range of 0.11 to 0.16. The feature group-related medians, except the cepstral-related ones, are similar in the particular experimental condition but differ across all conditions. The loudness-, LPC-, waveform-, and pitch-related medians provide a large gradient across the "Cabin" (or "Hall") and "Cabin MP" (or "Hall MP") condition, compared to the cepstral-related medians across the mentioned conditions. We conclude

that the "MP" condition affects cepstral features less than the other emobase features. Regarding only both "MP" conditions, the feature-related medians are similar, in contrast to the ones across the "Cabin" and "Hall" condition.

Analysis of RIR's Influence on Emotion Recognition: Table 2 depicts the results for the different cross-variant experiments. Interestingly the recognition results within the speaker cabin are significantly better than the corresponding results within the absorber hall ("Hall MP": F=12.3050, p=0.0025 and original F=37.8488, p=0.0000). In comparison to the clean EMO-DB, all recognition results are significantly decreased (p<0.0000).

Table 2: Emotion recognition performance (F_1 score) per experimental condition. The baseline recognition performance (training/ test on EMO-DB) is also given.

Experimental condition	F_1 (std) [%]
EMO-DB (Baseline)	78.18 (0.631)
"Cabin"	73.63 (0.104)
"Cabin MP"	75.69 (0.789)
"Hall"	72.02 (0.821)
"Hall MP"	74.57 (0.630)

Discussion

The results of our feature analysis go along with the emotion recognition results regarding all experimental conditions. We obtain the best feature and recognition performance results in the "MP" condition, where temporal effects of the room acoustics were suspended. The room acoustics analysed in this work have an impact on the emotion recognition performance up to $\Delta F_1 = 6\%$. EMO-DB in the "Cabin" condition provides a few more massive outliers than in the other conditions. We suppose that this is due to the cabin's room modes, which distort EMO-DB, especially in the low-frequency range.

Assuming that the "MP" versions of the RIRs contained the spectral properties of the investigated rooms mainly, and the measured RIRs contained spectral as well as temporal properties such as delay and reverberation, we conclude that both aspects play a role in emotion recognition. For the two rooms, that were investigated in this study, both the spectral and temporal properties caused a nearly equally strong impairment of emotion recognition. To what extent these results also apply for living rooms or office rooms is not clear. Possibly, due to the typically longer reverberation times in these rooms, temporal effects might be more dominant.

So far, we cannot explain why the feature and recognition performance in the "Hall" condition is worse than in the "Cabin" condition. We have not identified any significant mismatch in the measuring of the RIR or generating the artificial EMO-DB data for training and testing. However, we will focus it on further investigations.

Acknowledgments

The authors want to thank the Chair of Electromagnetic Compatibility of the Otto von Guericke University, which made it possible to use the absorber hall for the analyses presented in this paper. Special thanks go to Mathias Magdowski for the support during the recordings.

References

[1] B. Schuller et al. Emotion recognition in the noise applying large acoustic feature sets. *Proc. Speech Prosody 2006, Dresden*. 2006

[2] A. Tawari and M. M. Trivedi. Speech Emotion Analysis in Noisy Real-World Environment. *2010 20th International Conference on Pattern Recognition*. Aug. 2010, 4605–4608

[3] J. Höbel-Müller et al. Analysis of the influence of different room acoustics on acoustic emotion features. *Elektronische Sprachsignalverarbeitung 2019. Tagungsband der 30. Konferenz*. Dresden, Germany, 2019, 156–163

[4] J. Höbel-Müller et al. Analysis of the influence of different room acoustics on acoustic emotion features and emotion recognition performance. *Tagungsband - DAGA 2019*. Rostock, Germany, 2019, 886–889

[5] M. Gottschalk et al. Filtering-based analysis of spectral and temporal effects of room modes on low-level descriptors of emotionally coloured speech. *Elektronische Sprachsignalverarbeitung 2020. Tagungsband der 31. Konferenz*. Magdeburg, Germany, 2020, 219–226

[6] F. Burkhardt et al. A Database of German Emotional Speech. *Proc. of the Interspeech-2005*. Lissabon, Portugal, 2005, 1517–1520

[7] F. Eyben, M. Wöllmer, and B. Schuller. openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor. *Proc. of the 18th ACM International Conference on Multimedia*. MM '10. Firenze: ACM, 2010, 1459–1462

[8] P. Sprent and N. Smeeton. *Applied Nonparametric Statistical Methods*. Chapman and Hall/CRC, 2001

[9] J. Lefter et al. Cross-corpus analysis for acoustic recognition of negative interactions. *Proc. of the 6th ACII*. Xian, China, 2015, 132–138

[10] R. Böck et al. Comparative Study on Normalisation in Emotion Recognition from Speech. *Proc of the 9th IHCI 2017*. Ed. by Patrick Horain, Catherine Achard, and Malik Mallem. Cham: Springer International Publishing, 2017, 189–201

[11] M. Hall et al. The WEKA Data Mining Software: An Update. SIGKDD Explor. Newsl. 11.1 (2009), 10–18