Analysis of the influence of different room acoustics on acoustic emotion features and emotion recognition performance

Juliane Höbel-Müller[1], Ingo Siegert[2], Ralph Heinemann[1], Alicia Flores Requardt[1], Michael Tornow[1], Andreas Wendemuth[1]

[1]Cognitive Systems Group, Institute for Information and Communications Engineering

[2]Mobile Dialog Systems, Institute for Information and Communications Engineering

Email: juliane.hoebel@ovgu.de

Otto-von-Guericke University Magdeburg, 39016 Magdeburg, Germany

Introduction

Distant voice-based human-machine interaction (HMI) can be exposed to varying environmental conditions. The influence of an increasing speaker-to-microphone distance (SMD) that reduces the signal-to-noise ratio of speech and noise [1], in addition to changing speech directions induce varying acoustic effects onto the captured signal. Consequently, acoustic features are degraded. Furthermore, the performance of speech emotion recognition is compromised by previously unseen conditions, which is typically due to a mismatch between a recognition system's training and testing distribution. Therefore, a lot of effort is put in the matching of training conditions and test conditions. Remarkable progress has been made in distant-speech-recognition (DSR) [6, 7]. However, transferring knowledge from DSR to distant-speech-emotion-recognition (DSER) is only moderately possible.

So far, distant speech emotion recognition has been analysed in terms of superposed noise [10, 11], robust feature sets [12, 13] or feature pooling [9]. The impact of room acoustic characteristics on specific feature types and the performance of speaker state classification has been analysed [1, 4, 5]. In [4] the impact of reverberation using public room impulse responses for convolution of emotionally coloured speech is shown. A more detailed look at feature degradation is given by Eyben et al. [5], highlighting the decrease of the relative important energy-related features when introducing reverberation and noise. It is not possible to gain insights into the impact of room acoustics on the emotion recognition performance in isolation of specific acoustic features belonging to a special feature type. This issue is attributed in this paper.

This work is structured as follows: The recording setup is introduced including the description of Berlin Emotional Speech database (EMO-DB) and the microphone-loudspeaker configuration. Room acoustic characteristics of four real-life rooms including an anechoic chamber are determined and interpreted, before a feature analysis is conducted and emotion recognition experiments are performed.

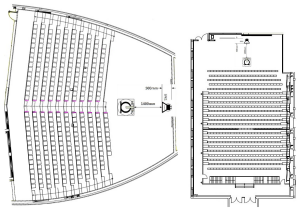
Experimental Design

EMO-DB utterances were replayed using a loudspeaker and re-recorded. The experimental design has been already described in detail in [27]. The benchmark speech database used is the studio-recorded Berlin Database of Emotional Speech (EMO-DB) [14]. Five female and five male professional actors spoke ten German usually emotionally neutral sentences. The final phrases were selected using a perception test, resulting in 494 utterances in the following seven emotion categories: anger, boredom, disgust, fear, joy, neutral and sadness. The original recordings sampled at 16 kHz provide a high audio quality, minimizing extrinsic variability factors.

In order to compare the impact of room acoustics on re-recordings, the following equipment was used: a) Sennheiser ME66 microphone, b) Yamaha 01V96i audio interface, c) Neumann KH120A loudspeaker and d) Cubase AI6 as recording software. EMO-DB was re-recorded at a sampling rate of 44.1 kHz in acoustically different rooms, all located at the Otto-von-Guericke University in Magdeburg. A reference and four test rooms were chosen to cover various reverberated indoor environments, see Figure 1. By re-recording in an anechoic chamber, a reference was created. Subsequent re-recordings were done in an Ambient Assisted Living (AAL) room, a modern lecture hall (LH), an old LH and a staircase. The AAL room is comparable to a living room - a narrow room with a carpet, fully furnished. The modern and old LH are almost equal in volume. In contrast to the old LH, the modern one is equipped with state-of-the-art wall absorbers, in contrast to the old one. The staircase does not include any acoustic wall treatments.

The DSER is influenced by the SMD, room acoustics effects and the signal-to-noise ratio (SNR) of the speech and ambient noise [1]. Two very contrary positions for the recording experiments were selected. A *baseline* position is defined by a speaker-microphone distance of 1.4 m in an azimuth angle of 45°. Using this method, cross-room comparable re-recordings were created. A *distant* position was defined as the prolongation of the *baseline* SMD-axis towards the very far end possible in the specific room. Room impulse responses $h(t)$ were obtained in 44.1 kHz using a sine wave sweep. The acoustic characteristics of these were provided by six third octave bands in the range of 125 to 4000 Hz. Regarding DIN EN ISO 3382 [3], two pairs of objective and subjective room acoustics are determined: (1) the Clarity (C50), the Definition index (D50), (2) the Reverberation Time (T30)

(a) View to (l.) and from the loudspeaker (r.) in the AAL room (59 m^3).



(b) Ground plan of the modern (l.) and the older lecture hall (r.), with 1990 m^3 and 1922 m^3.

(c) Baseline setup (l.) in the staircase and a top view of it (r). The room volume is 134 m^3.

Figure 1: Illustration of analyzed rooms.

and the Early Decay Time (EDT).

As this experiment only aimed at the investigation of room acoustic characteristics on speech features and emotion recognition performance in speech, effects in the re-recorded EMO-DB due to a low SNR had to be mitigated. A SNR-optimal pair of power values (dB) of the source and reverberated signal was determined experimentally in the anechoic chamber. This pair of power values was conveyed to the *baseline* re-recording in the other rooms.

Determination of Room Acoustics

Figure 2 depicts the measured acoustic characteristics of the considered rooms. The boxplots for each third octave band gives an impression of the characteristics' overall distribution, by measuring room acoustics at several places. The ✖ indicates the value of the characteristic for the *baseline* re-recording. The ○ indicates the value of the characteristic for the *distant* re-recording. Table 1 gives an overview of average mid-frequency values of C50, D50, T30 and EDT.

According to DIN 18041 [2], the determined C50 and D50 values in the staircase were approaching the corresponding limit for good speech intelligibility, e.g. C50 = 0 dB or D50 = 50%. The room impulse response (RIR) varied not only in terms of descending speech clarity for the wider rooms, but also in terms of higher reverberation times T30 and EDT, as was expected. Yet, every room, except the staircase, fulfilled the volume-dependent nominal values for average T30 regarding speech performances [19].

Statistical Feature Analysis

The OpenSMILE toolkit was utilized [8] for emobase feature extraction. The feature set emobase comprises of 52 features: 26 acoustic low-level descriptors (LLDs) related to energy, pitch, spectral, cepstral, mel-frequency and voice quality and corresponding first order regression coefficients (Deltas). The derived functionals were skipped in this investigation. The 26 LLDs were computed from each of the 494 re-recorded utterances, resulting in 494 utterances×(4+1 rooms)×2 positions = 4940 utterances. Feature values were extracted on frame-level, i.e., there were 52 feature value sets per utterance, each comprised measures of 25 ms windows of the utterance. Pairs of feature value sets were created, whereby the first element of each pair corresponded to the anechoic chamber and the second element to a *baseline* or *distant* position in a reverberated room.

By applying paired t-tests, the means of the two feature value sets are compared in terms of their differences.The staircase re-recording offered the main number of highly significant different samples with medium to large effect sizes, followed by the old and modern LH, and the AAL room. Analysing feature degradation in general only LLDs were impacted by room acoustics and not their Deltas. Especially, the lower MFCCs mfcc_sma[1-4] were impacted, which are related to 50 % of the highly significant samples. mfcc_sma[2] is the most impacted LLD over all emotions and rooms. An increasing number of feature value sets related to MFCCs were degraded from *baseline* to *distant* condition. For further references on the statistical feature analysis regard our previous paper [27].

Influence on Emotion Recognition Performance

State-of-the-art emotion recognition experiments comparable to [24] were conducted. In contrast to those, a Leave-One-Speaker-Out (LOSO) validation scheme was chosen to better represent realistic applications. For feature extraction, the same feature set as for the statistical analyses had been used, also including their functionals, resulting in 988 features. Afterwards, standardisation as a normalisation technique is used to eliminate differences between the data samples [25]. Recognition was conducted using SVMs with linear kernel and a cost factor of 1 utilizing WEKA [26]. As performance measure, the unweighted average recall (UAR), unweighted average precision (UAP) and the F-measure (F_1) were calculated. The results are given in Table 2. In the *baseline*

Table 1: Baseline measurement: Averaged values of C50, D50, T30 and EDT for third octave bands in the range of 125 Hz to 4 kHz.

Room	C50 [dB]	D50 [%]	T30 [s]	EDT [s]
AAL room	10.91	87.55	0.38	0.37
Modern LH	10.55	85.71	0.83	0.14
Old LH	9.13	81.46	1.37	0.31
Staircase	2.30	62.23	1.74	1.56

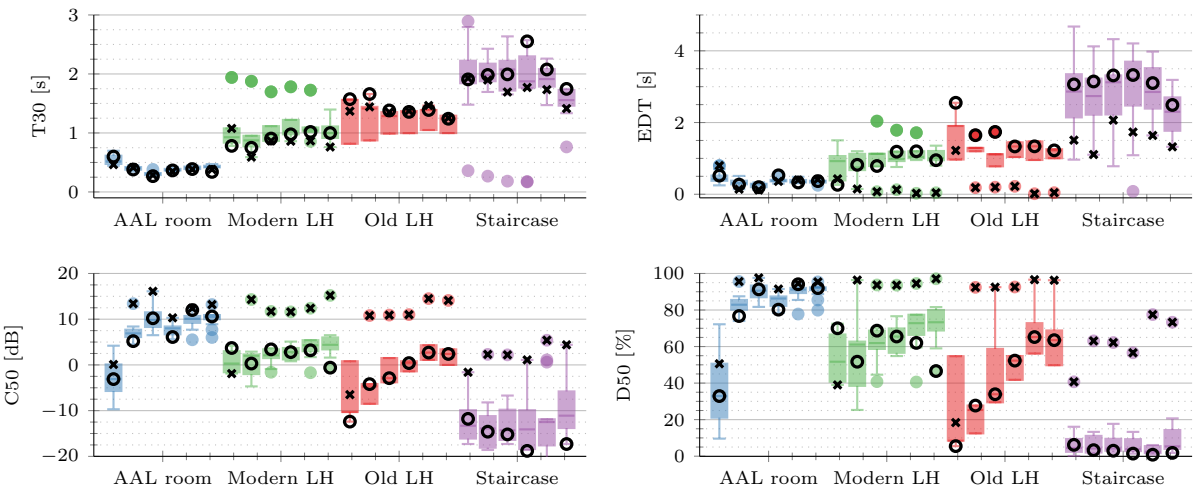


Figure 2: Boxplots showing *baseline* (✖), *distant* (○) and other exposed acoustic measurements per room: T30, EDT, C50 and D50 are measured in six third octave bands in the range of 125 to 4000 Hz.

Table 2: DSER performances for *baseline* and *distant* EMO-DB re-recordings. The best or lowest ones are highlighted in **bold** or *italics*.

Room	Re-recording	UAR	UAP	F_1
Anechoic chamber	Baseline	**0.80**	**0.78**	**0.76**
AAL room	Baseline	**0.75**	**0.74**	**0.71**
Modern LH	Baseline	0.73	0.73	0.70
Old LH	Baseline	0.72	0.73	0.70
Staircase	Baseline	*0.71*	*0.71*	*0.68*
Anechoic chamber	Distant	-	-	-
AAL room	Distant	0.73	0.72	0.70
Modern LH	Distant	0.68	0.69	0.65
Old LH	Distant	0.63	0.63	0.60
Staircase	Distant	*0.40*	*0.46*	*0.39*

condition, the AAL room provided the best DSER performance, followed by the modern and old LH, and last the staircase. The decreasing *baseline* DSER performances in the AAL room, modern and old LH and staircase correlate with the D50 values in the mentioned rooms (Spearman's rho $r_s = 1$). Increasing T30 or EDT values are associated with decreasing DSER performances ($r_s = -1$). Compared to T30 or EDT, D50 may a more intuitive measure to predict DSER performance in this experiment. Except for the staircase, the DSER performance from *baseline* to *distant* condition is decreasing similarly in each room (cf. [1]). Highly reflective and a lack of absorbing surfaces in the staircase are the reasons for the highest performance drop from *baseline* to *distant* condition.

Conclusion

In order to enable comparisons of existing studies a well known emotion speech database was re-recorded in four different prototypical rooms and the room acoustic variations in reverberated indoor environments were analysed. The rooms were acoustically characterised by several acoustic measures. To cover different acoustics for each room including a staircase, two different locations were examined for their feature degradation and emotion recognition performance.

The reported results present fundamental knowledge towards dealing with acoustic effects in reverberated rooms in speech applications. Most of the degraded features were determined for in the staircase, which highly differed from both LHs in terms of C50, D50, T30 and EDT. Highly impacted features are the lower MFCCs, which accounted for approximate 50% of the degradation. The highest degradation over the different room acoustics and emotions were observed for mfcc_sma[2]. Least impacted are all of the LLDs' Deltas. These results were also present in the DSER, as the performance decreased over all rooms in a nearly similar amount, except for the highly reverberated staircase. In detailed analysis of the relation of acoustic measures and recognition performance revealed that D50 is an adequate measure to estimate the decrease in DSER.

Further evaluation will be conducted by using a larger feature set with more LLDs, such as the emolarge feature set, which is also common in the field of speaker state recognition (cf. [4]). Additionally, correlation coefficients of degraded samples with the corresponding clean sample will be ascertained in order to examine the temporal behaviour of features in an utterance.

References

[1] Ahmed, M. Y. Z. Chen, E. Fass, J. A. Stankovic: Real time distant speech emotion recognition in indoor environments. Proc. MobiQuitous, Melbourne, Australia. 215–224, 2017.

[2] Nocke, C.: Die neue DIN 18041 – Hörsamkeit in Räumen. Lärmbekämpfung Bd 11, 2016.

[3] DIN EN ISO 3382: Akustik–Messung von Parametern der Raumakustik, 2000.

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

[5] Eyben, F. F. Weninger, B. Schuller: Affect recognition in real-life acoustic conditions – a new perspective on feature selection. Proc. of the Interspeech-2013, Lyon, France. 2044–2048, 2013.

[6] Hsu, W.-N. J. Glass: Extracting Domain Invariant Features by Unsupervised Learning for Robust Automatic Speech Recognition. IEEE ICASSP, 5614–5618, 2018.

[7] Morgan, N.: Deep and wide: Multiple layers in automatic speech recognition. IEEE Transactions on Audio, Speech, and Language Processing, 20(1), 7–13, 2012.

[8] Eyben, F. M. Wöllmer B. Schuller: openSMILE – the munich versatile and fast open-source audio feature extractor. Proc. of the 18th ACM International Conference on Multimedia,1459–1462. 2010.

[9] Avila, A. R. Z. A. Momin, J. F. Santos, D. O'Shaughnessy, T. H. Falk: Feature pooling of modulation spectrum features for improved speech emotion recognition in the wild. IEEE Transactions on Affective Computing, 1–1, 2018.

[10] Schuller, B. D. Arsic, F. Wallhoff, G. Rigoll: Emotion recognition in the noise applying large acoustic feature sets. Proc. Speech Prosody, Dresden. s.p., 2006.

[11] Tawari, A. M. M. Trivedi: Speech emotion analysis in noisy real-world environment. 20th International Conference on Pattern Recognition, 4605–4608. 2010.

[12] Kim, E. H. K. H. Hyun, Y. K. Kwak: Robust emotion recognition feature, frequency range of meaningful signal. ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication., 667–671. 2005.

[13] Lee, K.-K. Y.-H. Cho, K.-S. Park: Robust feature extraction for mobile-based speech emotion recognition system. , In: G. Irwin, D. Huang (eds.): Intelligent Computing in Signal Processing and Pattern Recognition, 470–477. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[14] Burkhardt, F. A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss: A database of german emotional speech. Proc. of the Interspeech-2005, 1517–1520. Lissabon, Portugal, 2005.

[15] Bradley, J. S. R. D. Reich, S. G. Norcross: On the combined effects of signal-to-noise ratio and room acoustics on speech intelligibility. The Journal of the Acoustical Society of America, 106(4), 1820–1828, 1999a.

[16] Ahnert, W.: Einsatz elektroakustischer Hilfsmittel zur Räumlichkeitssteigerung, Schallverstärkung und Vermeidung der akustischen Rückkopplung. Dissertation, Technische Universität Dresden, 1975.

[17] Bradley, J. R. Reich, S. Norcross: A just noticeable difference in c50 for speech. Applied Acoustics, 58(2), 99–108, 1999b.

[18] Jordan, V. L.: Acoustical design of concert halls and theatres: a personal account. Elsevier Applied Science, London, 1980.

[19] Weinzierl, S.: Handbuch der Audiotechnik. Springer Science and Business Media, Berlin, 2008.

[20] Höhne, R. G. Schroth: Zur Wahrnehmbarkeit von Deutlichkeits-und Durchsichtigkeitsunterschieden in Zuhörersälen. Acta Acustica united with Acustica, 81(4), 309–319, 1995.

[21] Dunn, O. J.: Multiple comparisons among means. Journal of the American Statistical Association, 56(293), 52–64, 1961.

[22] Cohen, J.: A power primer. Psychological bulletin, 112(1), 155, 1992.

[23] Naylor, P. A. N. D. Gaubitch: Speech Dereverberation. Springer Publishing Company, Incorporated, 1st ed. , 2010.

[24] Lefter, I., Nefs, H.T., Jonker, C.M., Rothkrantz, L.: Cross-corpus analysis for acoustic recognition of negative interactions. In: Proc. of the 6th ACII. 132–138. Xian, China (2015).

[25] Böck, R. O. Egorow, I. Siegert, A. Wendemuth: Comparative Study on Normalisation in Emotion Recognition from Speech, In: P. Horain, C. Achard, M. Mallem (eds.): Intelligent Human Computer Interaction, 189–201. Springer International Publishing, Cham (2017).

[26] Hall, M. Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. SIGKDD Explor. Newsl. 11(1), 10–18 (2009).

[27] Höbel-Müller, J. I. Siegert, R. Heinemann, A. F. Requardt, M. Tornow, A. Wendemuth: Analysis of the influence of different room acoustics on acoustic emotion features. In: Elektronische Sprachsignalverarbeitung 2019: Tagungsband der 30. Konferenz, Dresden, March 2019, TUDpress, 2019, 156–163 (2019).