ANALYSIS OF THE INFLUENCE OF DIFFERENT ROOM ACOUSTICS ON ACOUSTIC EMOTION FEATURES

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*
Otto-von-Guericke University Magdeburg, 39016 Magdeburg, Germany
juliane.hoebel@ovgu.de

Abstract: In automatic analyses of speech and emotion recognition, it has to be ensured that training and test conditions are similar. The presented study aims to investigate the influence of certain room acoustics on common features used for emotion recognition. As a benchmark database this study focuses on the Berlin Database of Emotional Speech. The following rooms were analysed: a) modern lecture hall, b) older lecture hall, and c) staircase. For all rooms and their different recording setups, different acoustic measures were captured. The speech recordings analysed in this paper were realized only at the ideal locations within the rooms. Afterwards, 52 features (LLDs of emobase) were automatically extracted using OpenSMILE and a sample-wise statistical analysis (paired t-test) was carried out. Therefore, the number of acoustically degraded features and its effect size can be linked to the acoustic parameters of the different recording experiments. As result, 15% of the degraded samples show a highly significant difference regarding all considered rooms. Especially MFCCs account for approximate 50% of the degradation. Furthermore, the degradation is analysed depending on the emotion and room acoustic.

1 Introduction

Voice-based human-machine interaction (HMI) "in the wild" is exposed to varying environment conditions. Increasing speaker-to-microphone distance reduces the signal-to-noise ratio of the speech and noise [1]. In addition, changing speech directions induce varying acoustics effects into the captured signal. Consequently, acoustic features are degraded (cf. [2, 3]).

Also, the performance of speech emotion recognition can be compromised by previously unseen conditions, which is typically due to a mismatch between a recognition system's training and testing distributions. Research areas in HMI, investigating the size of the difference between ideal training and distorted testing distribution, are referred to as distant-speech-emotion-recognition (DSER) and distant-speech-recognition (DSR). While remarkable progress has been made in DSR [4, 5], it is moderately possible to transfer knowledge from DSR to DSER. Speech emotion recognition is usually based on large feature sets, which contain low-level descriptors (LLDs), delta regression coefficients, and their functionals [6], whereas automatic speech recognition needs phonemes and language models and typically a limited number of features such as Mel Frequency Cepstral Coefficients (MFCCs) and corresponding delta regression coefficients (Deltas). Commonly used classifiers are support vector machines (SVMs), Gaussian mixture models (GMMs) and random forests (RFs), which are applied for speech emotion recognition, whereas hidden Markov models (HMMs) are appropriate for automatic

speech recognition. In both research areas, the current trend is applying deep learning techniques [5, 7].

So far, speech emotion recognition "in the wild" has been analysed in terms of superposed noise [8, 9], robust feature sets [10, 11] or feature pooling [7]. Furthermore, there are studies showing the impact of room acoustic characteristics on feature types and performance of speaker state classification [1, 2, 3]. Schuller et al. [2] show the impact of reverberation using public room impulse responses for convolution of emotional coloured speech. The authors report the suitability of feature types regarding different room impulse responses. In order to do this, the accuracies of classification systems are compared [2]. A more detailed look at the features is given by Eyben et al. [3]. The features are sorted in terms of degradation induced by both reverberation and background noise. Especially, the relative importance of energy-related features decreases when introducing reverberation and noise [3]. However, it is not possible to gain insight into the impact of room acoustics in isolation on specific acoustic features belonging to a special feature type. This issue is attributed in this paper.

In order to conduct a statistical feature analysis, emotionally coloured speech from a high quality benchmark corpus is re-recorded in various rooms. The selected real-life rooms are acoustically different to cover reverberated indoor environments, whereby an anechoic chamber provides a reference for the comparison. Our contribution is providing feature sets least and most impacted by room acoustics characteristics. Features analysed are LLDs and first order regression coefficients (Deltas) representing a subset of the emobase feature set. By applying paired t-tests, the means of each reverberated feature value set and the corresponding clean feature value set related to an anechoic chamber, are compared in order to verify how significant the differences are. A re-recording setup and a first insight regarding the aforementioned feature analysis is given.

In the remainder of this paper, first, our recording setup will be introduced in Section 2 including the description of Berlin Emotional Speech database (EMO-DB), the microphone-loudspeaker configuration and the understanding of selected signal processing fundamentals. Room acoustics characteristics of four real-life rooms including an anechoic chamber are determined and interpreted in Section 3. Statistical feature analysis and power analysis is conducted in Section 4 before concluding in Section 5.

2 Experimental Design

2.1 Emotional Speech Corpus

The benchmark speech database used in this study is the Berlin Database of Emotional Speech (EMO-DB) [12] containing ten professional actors as speakers (five female). The female speakers are on average 30.6 $\pm$ 5.6 years old and the male speakers are on average 28.8 $\pm$ 3.1 years old. Each of them simulates different emotion categories when asked to recite ten different German utterances with neutral semantic content. Overall, the database contains 494 different utterances spanning between 2-5 seconds 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.

2.2 Recording Setup

2.2.1 Hardware and Software

In order to compare the impact of room acoustics on re-recordings, a similar microphone (Sennheiser ME66), audio interface (Yamaha 01V96i), loudspeaker (Neumann KH120A) and

recording software (Cubase AI6) setup was used. This hardware setup is characterized by a highly linear frequency response. The ME66 is especially suitable for picking up quiet signals in noisy environments, as sound events outside the main speech direction are suppressed. It is characterized by a distinct directional characteristic, a high degree of bundling, a low inherent noise (10 dB according to DIN IEC 651), a reasonable maximum sound pressure level (126 dB) and a high sensitivity. The frequency range is between 40 Hz and 20 kHz ($\pm$ 2.5 dB).

2.2.2 Microphone-loudspeaker configuration

DSER is influenced by the microphone-speaker distance, room acoustics effects and the signal-to-noise ratio (SNR) of the speech and ambient noise [1]. Additionally, the SNR of the speech and noise is influenced by room acoustics characteristics [13]. By aligning the microphone and loudspeaker to each other with a similar distance of 1.4 m in an azimuth angle of 45°, cross-room comparable re-recordings are created. As this experiment only aims at the investigation of room acoustics characteristics on speech features and emotion recognition performance in speech, effects in the re-recorded EMO-DB due to a low SNR have to be mitigated. In order to achieve this, a SNR-optimal pair of power values (dB) of the source and reverberated signal is determined experimentally in the anechoic chamber. This pair of power values is conveyed to any other rooms.

2.2.3 Re-recording EMO-DB

EMO-DB was re-recorded with 44.1 kHz sampling rate in acoustically different rooms, which are located at the Otto-von-Guericke University in Magdeburg. In order to exclude overlap effects caused by previously played utterances, EMO-DB was played with a one-second pause after each utterance. Equation 1 represents the re-recorded speech signal $s(t)$ as a convolution of the source signal $x(t)$:

$$s(t) = x(t) * h(t) + n(t), \tag{1}$$

where $h(t)$ is the room impulse response (RIR) of the channel from the source to microphone and $n(t)$ is background noise in the room. The RIR describes the acoustic properties of a room in time domain in terms of sound propagation and reflections for a specific microphone-loudspeaker configuration. By convolving the EMO-DB utterances with different IRs, various room acoustic effects are created into the EMO-DB recordings.

3 Determination of Room Acoustics

3.1 Selected Room Acoustics

Four rooms were chosen to cover various reverberated indoor environments and to reach small to large reverberation times. By re-recording in (a) an anechoic chamber, a reference was created. Subsequent re-recordings were done in (b) a modern lecture hall, (c) an old lecture hall and (d) a staircase on the ground floor. The modern and old lecture hall are almost equal in volume, however, the modern lecture hall is equipped with state-of-the-art absorber walls, in contrast to the old one. The staircase does not include any acoustic treatments.

Room impulse responses $h(t)$ were obtained in 44.1 kHz using a maximum length sequence signal. Acoustics characteristics of these are provided for octave bands in the range of 125 to 4000 Hz using the Computer Aided Room Analyser (CARMA) Vers.$\tilde{4}$.0 software and a Behringer ECM8000 ultra-linear condenser microphone with omnidirectional pattern. Regarding DIN EN ISO 3382, two pairs of objective and subjective room acoustics are determined: (1) the Clarity (C50) and the Definition index (D50), representing (logarithmic) ratios between a

fraction and the entire or remaining RIR energy, and (2) the Reverberation Time (T30) and the Early Decay Time (EDT), which are obtained from the decay curve.

The early-to-late arriving sound energy ratio C50 is based on the assumption that the sound energy, which refers to a period of 50 ms after the arrival of direct sound, supports the clarity of speech as perceived by human ears [14]. Later parts would be detrimental to the clarity [14]. Accordingly, C50 is calculated as shown in equation 2.

$$\mathrm{C50} = 10 \cdot \lg \frac{\int_0^{50\ \mathrm{ms}} h^2(t)dt}{\int_{50\ \mathrm{ms}}^{\infty} h^2(t)dt}, \mathrm{dB}. \tag{2}$$

In order to determine early-to-total sound energy ratio D50, the energy portion is determined within the first 50 ms and related to the total energy, as described in equation 3.

$$\mathrm{D50} = \frac{\int_0^{50\ \mathrm{ms}} h^2(t)dt}{\int_0^{\infty} h^2(t)dt}, \%. \tag{3}$$

Compared to D50, C50 would be more related to subjective assessments of the clarity [15].

The reverberation time T30 is determined by measuring the sound pressure drop in a range from -5 dB to -35 dB. According to [16], the EDT is determined in a range between 0 dB and -10 dB. Compared to T30, EDT corresponds better to the subjectively perceived reverberation time. The reason for that is the initial portion of the sound decay, which is mainly responsible for subjective impression of reverberation in a room [17].

3.2 Measurement Determination

Table 1 – Room characteristics and averaged mid-frequency values of C50, D50, T30 and EDT for octave bands in the range of 125 to 4000 Hz.

Room	Volume [m^3]	C50 [dB]	D50 [%]	T30 [s]	EDT [s]
Anechoic chamber	19	30.10	90.61	0.13	0.15
Modern lecture hall	1990	10.55	85.71	0.83	0.14
Old lecture hall	1922	9.13	81.46	1.37	0.31
Staircase	134	2.30	62.23	1.74	1.56

Table 1 summarizes the volume and averaged acoustic characteristics of the considered rooms for the baseline microphone-loudspeaker configuration (see 2.2.2). According to DIN 18041, the C50 and D50 values have not fallen below the limit for good speech intelligibility, a value of C50 = 0 dB or D50 $\geq$ 50%. As one can see, the determined C50 and D50 values in the staircase are approaching the corresponding limit. The perception of clarity differences is limited by $\triangle$C50 $\approx \pm$2.5 dB [18]. Accordingly, the subjectively perceived clarity in both lecture halls resemble each other, but highly differs from the one in the staircase.

The RIR varies not only in terms of descending speech clarity for the wider rooms, but also in terms of higher reverberation times T30 and EDT, as one would expect. Yet, every room, except the staircase, fulfils the volume-dependent nominal values for average T30 regarding speech performances [17]. The relatively low EDT value can be explained by the fact that the acoustic measurement was carried out in the vicinity of an absorber wall.

By measuring room acoustics at several places, varying room impulse responses inside a room are obtained. As mentioned above, one acoustic measurement was conducted with a cross-room equal microphone-loudspeaker distance as described in Section 2.2.2, whereby the microphone was placed in the middle of the lecture halls and the staircase. This acoustic measurement

is referred to as baseline measurement, per room. Twelve other measurements were conducted at exposed places inside both lecture halls and the staircase. Then, frequency-dependent quantities of measured values were obtained. These quantities are depicted in Figure's 1 boxplots displaying variation in RIRs for the octave bands in the range of 125 to 4000Hz, which is most important for speech.

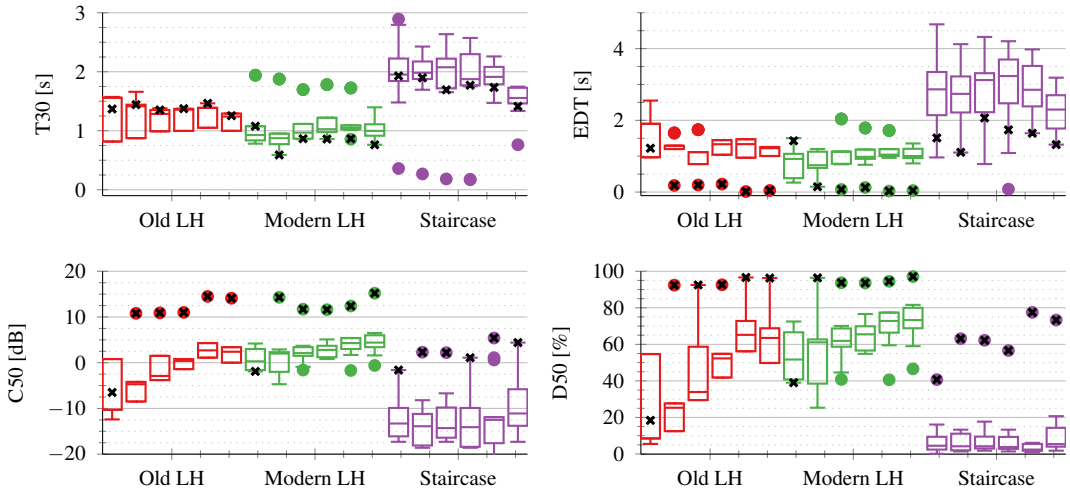

Figure 1 – Baseline (×) and other exposed acoustic measurements per room: C50, D50, T30 and EDT are measured in octave bands in the range of 125 to 4000 Hz. Selected rooms are an old lecture hall (LH), a modern LH and a staircase.

4 Statistical Feature Analysis

4.1 Drawing Samples

OpenSMILE feature extraction toolkit [6] is used to extract a subset of the emotion specific emobase feature set, namely 52 features as 26 acoustic LLDs related to energy, pitch, spectral, cepstral, mel-frequency and voice quality and corresponding first order regression coefficients (Deltas). The 26 LLDs are computed from each of the emotionally coloured utterances, which are re-recorded in the four rooms. Feature values are extracted on frame-level, i.e., there are 52 feature value sets per utterance, each comprising measures of each 25 ms windows of the utterance. 77064 pairs of feature value sets (494 utterances$\times$52 features$\times$3 rooms) are created, whereby the first element of each pair corresponds to the anechoic chamber and the second element to the modern lecture hall, the old one or the staircase. By applying paired t-tests, the means between the two feature value sets are compared in order to verify how significant the difference is. In the following, a feature value set corresponding to the modern lecture hall, the old one or the staircase is referred to as samples for reasons of simplicity.

4.2 Paired Difference Test Statistics

15% of the samples show a highly significant difference with medium to large effect sizes, regarding a Bonferroni-corrected [19] significance level (paired t-test, $p < 0.01/77064$) and an effect size of $|d| = 0.5$ defined by Cohen [20]. The LLDs are related to these highly significant different samples, whereas none of the Deltas is related to those. As one would expect, the staircase provides the main number of the highly significant different samples, namely 40%, while the old and modern lecture hall provide each roughly 30%. Average effect sizes avg($|d|$)

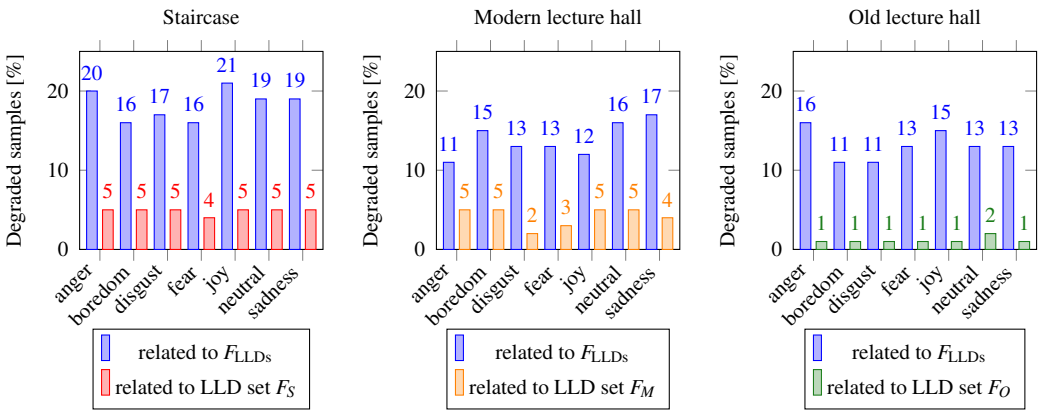

Figure 2 – Per room and emotion, percentage of highly significant different samples (blue), related to F_{LLDs}, and, per room, percentage of highly significant different samples, related to F_S, F_M or F_O (red, orange or green).

are determined to quantify the magnitude of the differences: $\text{avg}(|d|) = 0.73 \pm 0.2$, $\text{avg}(|d|) = 0.72 \pm 0.18$ and $\text{avg}(|d|) = 0.69 \pm 0.18$ are measured in the modern lecture hall, in the staircase, and in the old lecture hall, respectively.

Per room and emotion category, Figure 2 summarizes the percentages of highly significant different samples, which are related to $F_{\text{LLDs}} = \text{emobase} \setminus \{\text{Deltas}\}$ (blue bar). Then, LLD sets are determined containing LLDs, which cause approximate 50% of the degraded samples related to a specific room and emotion category. Initially, these LLD sets involve four to seven LLDs. Next, these sets are intersected per room to obtain new LLD sets sharing the same LLDs over all emotion categories per room. The resulting LLD sets are $F_M = \{\text{mfcc\_sma[1], mfcc\_sma[2], mfcc\_sma[3], mfcc\_sma[4]}\}$, related to the modern LH, $F_S = \{\text{lspFreq\_sma[5], mfcc\_sma[2], mfcc\_sma[4]}\}$, related to the staircase, and $F_O = \{\text{mfcc\_sma[2]}\}$, related to the old LH, whereby $|F_M| > |F_S| > |F_O|$. As can be seen, the lower MFCCs are overrepresented in F_M, F_S and F_O. Furthermore, exactly one equal LLD in the new LLD sets exists, namely mfcc_sma[2] with an average effect size $\text{avg}(|d|) = 0.71 \pm 0.2$. Figure 2 presents the percentages of highly significant different samples, which are only related to F_S, F_M and F_O, respectively, marked by the red, orange and green bars.

The rooms differ regarding the percentage of comparatively most and least degraded emotion categories, whereby "degraded" and synonyms are related to highly significant differences with medium to large effect sizes. In the staircase, the most degraded emotion category is joy, closely followed by anger. Contrarily, the emotion category sadness, closely followed by neutral, is most degraded in the modern lecture hall. In the old lecture hall, the most corrupted emotion category is anger, closely followed by joy. The least degraded emotion categories are boredom and disgust, respectively, in the old lecture hall, anger and joy in the modern one, and boredom and fear in the staircase.

Regarding the percentages of F_S, it becomes apparent that these numbers are evenly distributed across all emotion categories. The old lecture hall is comparable to this observation, unlike the modern lecture hall. As one could expect, the percentages of F_O are relatively small due to $|F_O| = 1$, whereas the cardinalities of F_M and F_S are $|F_M| = 4$ and $|F_S| = 3$.

5 Conclusion

This paper provides further insights on room acoustic variations in reverberated indoor environments. Furthermore, insights are presented regarding acoustic features impacted by different room acoustics. Such an analysis is of an existing importance when emotion recognition or related speaker state recognition in speech is applied in far-distant and voice-based HMI.

The reported results represent foundations towards dealing with acoustic effects in reverberated rooms in speech applications.

15% of the degraded feature value sets show a high impact, whereby most of the degraded features are determined in the staircase, which highly differs from both lecture halls in terms of C50, D50, T30 and EDT. Highly impacted features are the MFCCs, which acount for approximate 50% of the degradation. The highest degradation is observed for mfcc_sma[2] with an average effect size $\text{avg}(|d|) = 0.71 \pm 0.2$ over the different room acoustics and emotions. Intuitively, least impacted are all of the LLDs' Deltas.

Further analysis 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. [2]). Additionally, correlation coefficients of degraded samples with the corresponding clean sample will be determined in order to examine the temporal behaviour of features in an utterance. Additionally, the recognition performance loss in terms of various training and test conditions will be reported. Future work will also comprise strategies for de-reverberation (cf. [21]).

References

[1] AHMED, M. Y., Z. CHEN, E. FASS, and J. A. STANKOVIC: *Real time distant speech emotion recognition in indoor environments*. In *MobiQuitous*. 2017.

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

[3] 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.

[4] HSU, W.-N. and J. GLASS: *Extracting Domain Invariant Features by Unsupervised Learning for Robust Automatic Speech Recognition*. *IEEE ICASSP*, 2018.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] BRADLEY, J. S., R. D. REICH, and 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), pp. 1820–1828, 1999.

[14] AHNERT, W.: *Einsatz elektroakustischer Hilfsmittel zur Räumlichkeitssteigerung, Schallverstärkung und Vermeidung der akustischen Rückkopplung*. Ph.D. thesis, Technische Universität Dresden, 1975.

[15] BRADLEY, J., R. REICH, and S. NORCROSS: *A just noticeable difference in c50 for speech*. *Applied Acoustics*, 58(2), pp. 99 – 108, 1999.

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

[17] WEINZIERL, S.: *Handbuch der Audiotechnik*. Springer Science and Business Media, Berlin, 2008.

[18] HÖHNE, R. and G. SCHROTH: *Zur Wahrnehmbarkeit von Deutlichkeits-und Durchsichtigkeitsunterschieden in Zuhörersälen*. *Acta Acustica united with Acustica*, 81(4), pp. 309–319, 1995.

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

[20] COHEN, J.: *A power primer. Psychological bulletin*, 112(1), p. 155, 1992.

[21] NAYLOR, P. A. and N. D. GAUBITCH: *Speech Dereverberation*. Springer Publishing Company, Incorporated, 1st edn., 2010.