Towards a Better Understanding of the Effect of Reverberation on Speech Recognition Performance

Armin Sehr[1], *Emanuël A.P. Habets*[2], *Roland Maas*[1], and *Walter Kellermann*[1]

[1]Multimedia Communications and Signal Processing
University of Erlangen-Nuremberg
Cauerstr. 7, 91058 Erlangen, Germany
{sehr,maas,wk}@lnt.de

[2]Department of Electrical and Electronic Engineering
Imperial College London
Exhibition Road, London SW7 2AZ, United Kingdom
e.habets@imperial.ac.uk

***Abstract*—In order to tailor dereverberation approaches to automatic speech recognition (ASR) systems, it is important to thoroughly understand the effect of reverberation on ASR performance. In this work, the effect is analyzed by varying the shape of the room impulse response (RIR) using two design parameters that are useful for defining the target response of typical dereverberation algorithms. The parameters determine the amount of attenuation of the coefficients that correspond to reflections arriving with at least a delay of T after the direct-path component. By convolving clean speech signals with the modified RIRs, ideal late-reverberation suppression is simulated. By varying the level of attenuation A and the delay T, the influence of these design parameters on the recognition rates is investigated. Thus, guidelines for adjusting dereverberation algorithms to ASR systems are deduced.**

I. Introduction

Distant-talking microphone systems enable human/human and human/machine-interaction without tethering the user to a close-talking microphone. Due to the large distance between speaker and microphone in distant-talking scenarios, the microphone does not only pick up the desired signal but also background noise, interfering speakers, and the reverberation of the desired signal caused by multiple reflections of the sound waves at the boundaries of the enclosure. These interferences do not only reduce the perceived sound quality but also decrease the performance of ASR systems significantly [1]. The detrimental effect of reverberation on speech recognition rates can be explained by the dispersion of the feature vector sequences caused by reverberation, leading to overlap-masking of phonemes [2].

A promising way for achieving robust distant-talking ASR is to reduce these interferences by noise reduction, beamforming, and/or dereverberation algorithms before extracting the ASR features. Most signal enhancement algorithms are however designed to improve the signal characteristics, e.g., to maximize the signal-to-noise ratio, or to make the signal sound more pleasant. In other words, they are optimized for human listeners. It has turned out that speech enhancement schemes that are highly effective for improving the perceptual sound quality do not necessarily lead to a significant reduction of word error rates when applied as preprocessing units for ASR (see for example [1]). This observation indicates that the speech enhancement has to be adjusted to the ASR backend in order to be effective. A signal-based speech enhancement scheme aiming in this direction is described in [1]. By performing the enhancement directly in the feature domain, like in [3]–[5], the enhancement inherently focuses on the signal characteristics relevant for speech recognition.

In this paper, we focus on how signal-domain dereverberation algorithms can be adjusted to ASR. Dereverberation algorithms can be categorized into reverberation cancellation approaches and reverberation suppression approaches [6]. The algorithms in the first category aim at inverse filtering - either by first identifying the RIR and subsequently inverting it or by direct estimation of the inverse [7]–[9]. The algorithms in the second category aim at partial reverberation reduction and are generally robust to position changes of the source and receiver [6], [10], [11]. For reducing distortions of the speech signal, many dereverberation algorithms, e.g., [6], [9]–[12], do not try to reduce the early reflections, but only attenuate late reverberation. Such approaches appear to be particularly suitable for pre-processing in ASR, since straightforward normalization schemes, as, e.g., cepstral mean normalization (CMN) [13], can cope very well with the early reflections.

As a first step for tailoring such dereverberation approaches to ASR systems, we analyze the effect of reverberation on ASR performance. While several papers including [1], [14]–[16] have analyzed the dependencies of the recognition rate on basic parameters, like reverberation time T_{60} or the speaker-microphone distance d, we focus on parameters relevant for designing dereverberation algorithms. In particular, we investigate the relationship between the delay T where the reverberation suppression begins, the achieved level of attenuation A, and the recognition rate. Note that both T and A are control parameters of the algorithms proposed in [10] and [11].

The paper is structured as follows: Section II describes the method used for the analysis and Section III presents the detailed experimental conditions. The analysis results are discussed in Section IV and conclusions are drawn in Section V.

II. Analysis method

For our analysis, we simulate "ideal" late-reverberation suppression, where "ideal" implies that the speech signal is not distorted. Thus, an upper bound for the speech recognition performance that can be achieved with dereverberation approaches attenuating only the late reverberation is obtained. To this end, reverberant test signals were generated by convolving clean test signals with measured RIRs. The ideal late-reverberation suppression was achieved by attenuating all RIR coefficients that correspond to reflections arriving with at least a delay of T after the direct-path component by a certain level of attenuation A (in dB). Then the clean-speech data were convolved with such modified RIRs to obtain "dereverberated" test signals. The features extracted from these test signals were decoded by an ASR system so that recognition scores were obtained. In the following, this approach is described in more detail.

To generate reverberant test signals $x(n)$, clean speech signals $s(n)$ were convolved with a measured RIR $h(n)$ according to

$$x(n) = \sum_{m=0}^{N-1} h(m)\, s(n-m)\,, \tag{1}$$

where n is a sample index and N is the length of the RIR. Without loss of generality, we assume that the direct-path impulse corresponds to $n = 0$ in $h(n)$. Reverberation suppression algorithms such as the one described in [6], [10] aim at suppressing late reverberation caused

by reflections that arrive with at least a delay of T after the direct-path sound. These algorithms reduce overlap-masking, which is known to degrade speech intelligibility (see [6], [17]). To maintain a natural-sounding dereverberated signal and to avoid distorting the early speech component (defined as the direct-path signal plus reflections that arrive before T), the suppression of late reverberation is often limited to a certain level of attenuation A. In a similar way, an equalizer can be constructed to obtain the following target impulse response

$$\tilde{h}(n,T,A) = \begin{cases} h(n) & \text{for} \quad 0 \le n < T\, f_\mathrm{s}\ , \\ 10^{-\frac{A}{20\,\mathrm{dB}}}\ h(n) & \text{otherwise}\ , \end{cases} \tag{2}$$

where f_s is the sampling frequency. It is worthwhile noting that $\tilde{h}(n,T,0\,\mathrm{dB}) = h(n)$ for all T. Here we propose to use the target response $\tilde{h}(n,T,A)$ to study the recognition performance. Therefore, the obtained recognition results provide an upper bound for the recognizer when the aforementioned aim is achieved. The test data were generated by convolving $\tilde{h}(n,T,A)$ with the clean speech signal $s(n)$ according to

$$\tilde{x}(n,T,A) = \sum_{m=0}^{N-1} \tilde{h}(n,T,A)\ s(n-m)\ . \tag{3}$$

The feature vectors extracted from the signal $\tilde{x}(n,T,A)$ were fed into a recognizer with an acoustic model trained on clean data to determine the word accuracies for the respective test case. Given the total number of words N_W in the correct transcription, the number of word substitutions N_S, deletions N_D, and insertions N_I, the word accuracy is given by

$$\mathrm{ACC} = \frac{N_\mathrm{W} - N_\mathrm{D} - N_\mathrm{S} - N_\mathrm{I}}{N_\mathrm{W}} \cdot 100\,\%\ . \tag{4}$$

As a channel-based measure for reverberation, the "definition" D_{50} ("Deutlichkeit") [18] is used for explaining certain effects in the accuracy curves in the sequel. It is defined as the ratio of the energy of the direct sound plus early reflections arriving within the first 50 ms and the energy of the complete RIR, i.e.,

$$D_{50}(T,A) = \frac{\sum_{n=0}^{N_{50}-1} \tilde{h}^2(n,T,A)}{\sum_{n=0}^{N-1} \tilde{h}^2(n,T,A)} \cdot 100\,\%\ , \tag{5}$$

where $N_{50} = 50\,\mathrm{ms} \cdot f_\mathrm{s}$ is the number of samples corresponding to the first 50 ms.

III. Analysis Conditions

For evaluating ASR performance, a connected-digit recognition task was chosen. Because connected digit recognition does not require a language model, its recognition rate is solely determined by the degree the test data match the acoustic model. Since Hidden Markov Models (HMMs) trained on clean speech are employed as acoustic models, the recognition reflects how well the "dereverberated" test data match the clean-speech HMM. Here we employed a continuous-density recognizer with word-level HMMs using three Gaussian mixture components per state as reference recognizer. A more detailed description of this system, implemented with HTK [19], can be found in [20]. As speech features, 12 Mel-Frequency Cepstral Coefficients (MFCCs) including the 0-th coefficient plus the corresponding delta coefficients, extracted from the speech signal sampled at 20 kHz, were used. An analysis window length of 25 ms and a frame shift of 10 ms was used for the feature extraction. To calculate the delta coefficients, two preceding and two succeeding frames were taken into account. The experiments were performed using RIRs measured in three different rooms, which were selected to cover a wide range of reverberation conditions. The characteristics of these rooms are summarized in Table I, and further details are given in [20]. Note that the RIR for room B is taken from the sound scene database of the Real World Computing Partnership [21].

TABLE I
Summary of room characteristics: T_{60} is the reverberation time, N/fs is the length of the baseline RIR, d is the distance between speaker and microphone, D_{50} is the definition for the baseline RIR.

	Room		
	A	B	C
T_{60}	300 ms	780 ms	900 ms
N/f_s	200 ms	500 ms	700 ms
d	2.0 m	2.0 m	4.0 m
D_{50}	97.0 %	81.4 %	74.9 %

The TI digits corpus [22] was used both for test and training. A subset of the TI digits training set with 4579 connected digit utterances, corresponding to 1.5 hours of speech, was used for training. Continuous Gaussian mixture-density HMMs were used as acoustic models. A 16-state word-level HMM with no skips over states was used for each of the 11 digits ('zero' to 'nine' and 'oh'). Additionally, a three-state silence model with a backward skip from state three to state one was used. Mixtures with 3 Gaussian components were used as output densities for all HMMs.

A subset of 512 utterances from the TI digits test set (disjoint from the training data), corresponding to approximately 16 minutes of speech, was used for test. To obtain the dereverberated test data, the clean data were convolved with the modified RIRs $\tilde{h}(n,T,A)$ of the corresponding rooms for selected values of T and A. By feeding the MFCC features calculated from the "dereverberated" test signals $\tilde{x}(n,T,A)$ into the recognizer, the word accuracy $\mathrm{ACC}(T,A)$ as a function of T and A is obtained.

IV. Analysis Results

The upper row of Fig. 1 shows the word accuracy $\mathrm{ACC}(T,A)$ over T with A as parameter for all three rooms. The first 100 ms are enlarged in the center row. To assist the explanation of the word accuracy curves, the definition $D_{50}(T,A)$ is depicted in the lower row.

For a relatively low attenuation of $A = 5\,\mathrm{dB}$ or $A = 10\,\mathrm{dB}$ all accuracy curves show the same trend: The accuracy increases for growing T until a maximum is reached around 40 to 50 ms. Then, the accuracy decreases rapidly. Since we perform ideal late-reverberation suppression in the experiments, that is, the clean speech signal is not distorted regardless of T, this behavior appears to be unexpected. Actually, one would expect the accuracy to increase when T is reduced. The fact that the word accuracy decreases when T is reduced below some 40 or 50 ms can be explained as follows: For the given feature extraction parameters, the early reflections arriving before some 40 or 50 ms are obviously beneficial for the ASR system, similar to human perception (see, e.g., [18]). Only the reverberation arriving after this seems to be detrimental. As shown by the D_{50}-curves, for low attenuation, the ratio of the direct sound and early reflections energy to the total energy decreases when T is reduced below 50 ms, and thus the word accuracy decreases.

For higher attenuation of $A \ge 15\,\mathrm{dB}$, the accuracy curves are relatively flat within the initial 50 ms. That means, the word accuracy hardly changes if T is varied 10 to 50 ms. This behavior is also reflected by the relatively flat D_{50}-curves. It is worth noting that the

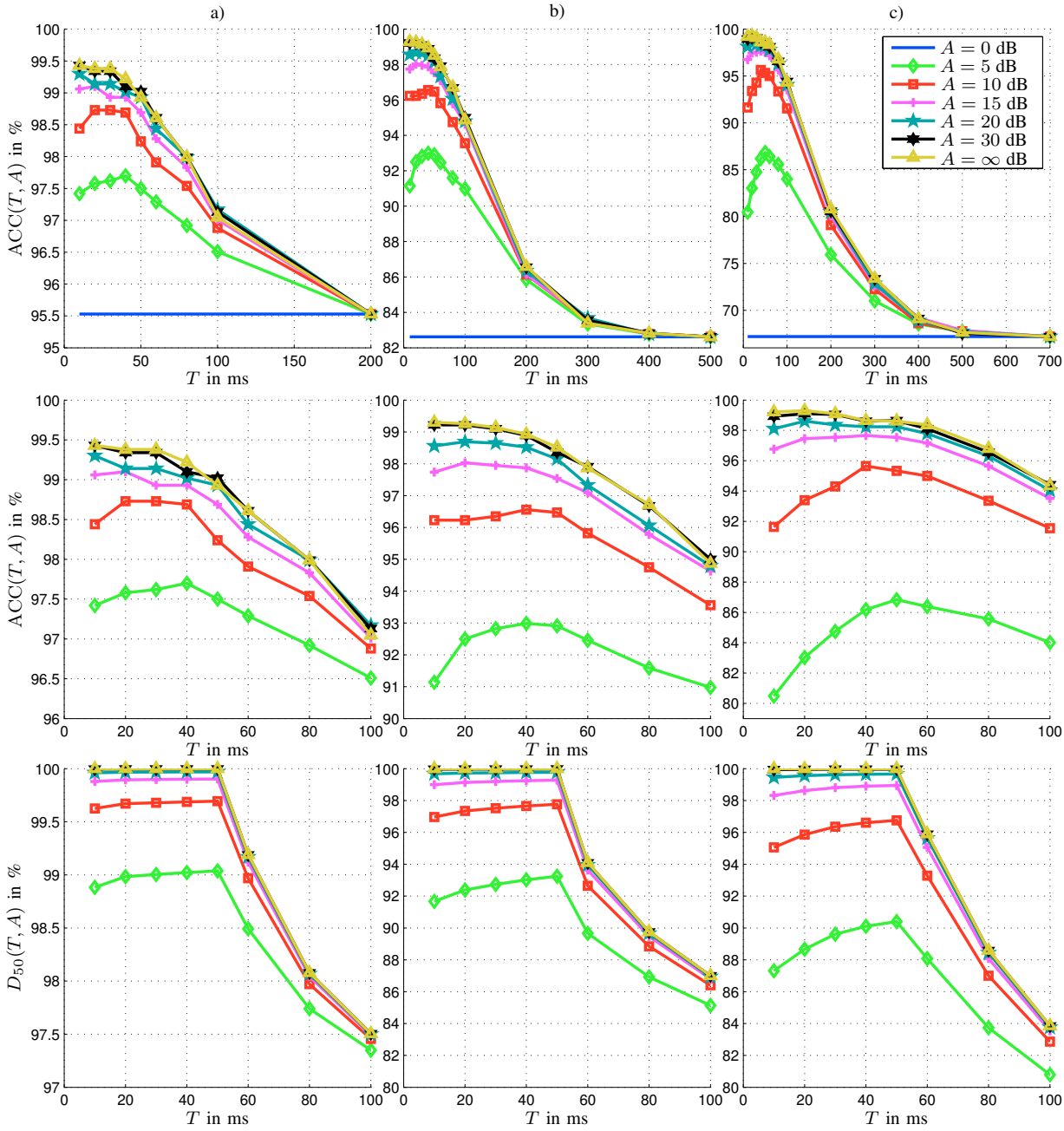

Fig. 1. Comparison of word accuracy ACC(T, A) and definition $D_{50}(T, A)$ for different rooms: a) Room A, b) Room B, c) Room C, upper row: ACC(T, A) over T with parameter A, center row: zoom of the upper row for $T \leq 100$ ms, lower row: $D_{50}(T, A)$ over T with parameter A.

improvement in recognition rate is only marginal if the attenuation is increased beyond $A = 15$ dB. While the accuracy curves show the same trends for all three rooms, there are some differences in the absolute numbers. First, the absolute word accuracies are decreasing with growing reverberation from Room A to Room C. For instance, with $T = 200$ ms and $A = 30$ dB, an accuracy of 95.5 % is achieved in Room A, while accuracies of only 86.2 % and 80.5 % are achieved in rooms B and C, respectively. This can be explained by the lower D_{50} in the more reverberant rooms B and C. Second, the time T yielding the best word accuracy tends to be slightly lower in Room A than in the other rooms.

From these results, some guidelines for designing dereverberation approaches for ASR can be deduced: It is not necessary to attenuate the early reflections arriving before $T = 50$ ms. If only a moderate late-reverberation attenuation can be achieved, reducing T below 50 ms would even be harmful to the ASR performance. Even if a very high attenuation can be achieved, attenuating only the reflections arriving after $T = 50$ ms still yields an ASR performance that is close to the optimum. Since increasing the attenuation beyond $A = 15$ dB does not yield large improvements in recognition rate, aiming for $A \approx 15$ dB instead of completely attenuating the late reverberation appears to be a sensible choice. Setting the design parameters to $T = 50$ ms and $A = 15$ dB in reverberation suppression can help to reduce distortion of the early speech components. Similarly, using $\tilde{h}(n, 50\,\text{ms}, 15\,\text{dB})$ as the target response of a reverberation cancellation algorithm (instead of a unit impulse response) can help

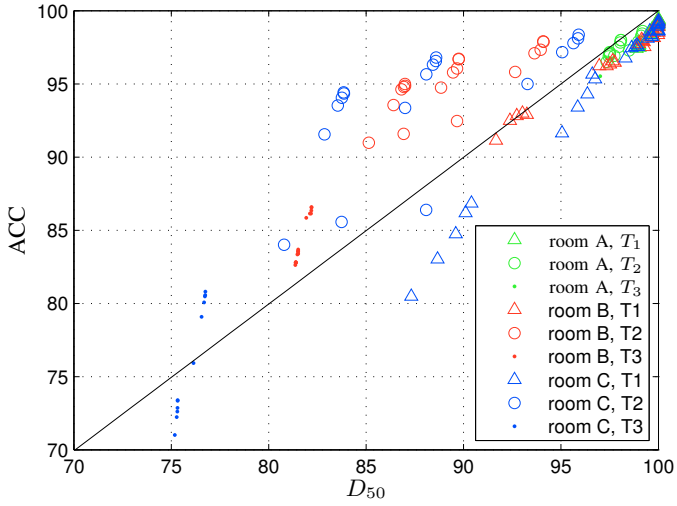

Fig. 2. Scatterplot ACC over D_{50}. T_1 means $T \leq 50\,\text{ms}$, T_2 means $50\,\text{ms} < T \leq 100\,\text{ms}$, T_3 means $T > 100\,\text{ms}$.

to make the inverse filtering more robust.

Comparing the accuracy curves and the D_{50}-curves in Figure 1, a close relationship between these two measures can be observed. To have a closer look at this relationship, Figure 2 shows the scatter plot of ACC over D_{50}. The fact that the data points in the scatter plot are aligned around a line and a relatively high correlation coefficient of 0.925 confirm that there is a strong dependency between ACC over D_{50}. This dependency can be described relatively accurately by a linear regression curve $\text{ACC} = \alpha D_{50} + \beta$, where $\alpha = 1.002 \approx 1.0$ and $\beta = -0.222 \approx 0$ for our data. These results indicate that the definition D_{50} could be used to predict the word accuracy for a given recognition task after the parameters α and β have been learned from a sufficiently large set of data points. Note however, that further research with a larger number of different rooms and a larger number of test conditions, including other parameter settings for feature extraction, is required to confirm these initial results.

V. Summary and Conclusions

The effect of reverberation on ASR performance was analyzed in this paper. To obtain an upper bound for the word accuracies achievable by late-reverberation suppression algorithms, clean-speech signals were convolved with room impulse responses where the coefficients corresponding to the reflections arriving with a delay greater than T after the direct sound are attenuated by a level A. Thus, ideal reverberation suppression not causing any distortions to the clean-speech signal was simulated. The results show that it is preferable to set $T = 50\,\text{ms}$ and to achieve an attenuation of $A \geq 15\,\text{dB}$. While $T = 50\,\text{ms}$ has been used in several dereverberation approaches with a perceptual justification, the experiments in this paper confirm that $T = 50\,\text{ms}$ is also a good choice for ASR. A further new insight of the study is that reducing T at the cost of a reduced attenuation leads to a decrease in ASR performance. This can be explained by the high correlation between the word accuracy and the definition D_{50}. Future work includes exploiting these relatively mild demands on the suppression of late reverberation for reducing distortions of the early speech components by reverberation suppression methods and for making inverse filtering methods more robust. Furthermore, the suitability of channel-based reverberation measures, like the definition D_{50}, for predicting the word accuracy of a given recognition task in reverberant environments will be further analyzed.

Acknowledgment

This work was partly supported by the Deutsche Forschungsgemeinschaft (DFG) under contract number KE 890/4-1 and by the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 226007 SCENIC.

References

[1] M. L. Seltzer, "Microphone array processing for robust speech recognition," Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, PA, USA, July 2003.

[2] A. Sehr and W. Kellermann, "Towards robust distant-talking automatic speech recognition in reverberant environments," in *Topics in Speech and Audio Processing in Adverse Environments*, E. Hänsler and G. Schmidt, Eds. Berlin: Springer, 2008, pp. 679–728.

[3] L. Deng, A. Acero, and X. Huang, "Large-vocabulary speech recogntion under adverse acoustic environments," *Proc. ICSLP*, pp. 806–809, 2000.

[4] M. Wölfel, "Enhanced speech features by single-channel joint compensation of noise and reverberation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 312–323, February 2009.

[5] A. Krueger and R. Haeb-Umbach, "Model based feature enhancement for automatic speech recognition in reverberant environments," *Proc. INTERSPEECH*, pp. 781–783, September 2009.

[6] E. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Technische Universiteit Eindhoven, 2007.

[7] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 2, pp. 145–152, February 1988.

[8] T. Nakatani, B. Juang, T. Yoshioka, K. Kinoshita, M. Delcroix, and M. Miyoshi, "Speech dereverberation based on maximum likelihood estimation with time-varying Gaussian source model," *IEEE Transactions on Audio, Speech, and Language Processing Processing*, vol. 16, no. 8, pp. 1512–1527, 2008.

[9] H. Buchner and W. Kellermann, "TRINICON for dereverberation of speech and audio signals," in *Speech Dereverberation*, P. Naylor and N. Gaubitch, Eds. Berlin: Springer, to appear.

[10] K. Lebart, J. Boucher, and P. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acoustica*, vol. 87, pp. 359–366, 2001.

[11] E. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 770–773, September 2009.

[12] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 534–545, May 2009.

[13] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *Journal of the Acoustical Society of America*, vol. 55, pp. 1304–1312, 1974.

[14] B. Kingsbury and N. Morgan, "Recognizing reverberant speech with RASTA-PLP," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1259–1262, 1997.

[15] A. Sehr, O. Gress, and W. Kellermann, "Synthetisches Multicondition-Training zur robusten Erkennung verhallter Sprache," *Proc. ITG Fachtagung Sprachkommunikation*, April 2006.

[16] R. Petrick, K. Lohde, M. Wolff, and R. Hoffmann, "The harming part of room acoustics in automatic speech recognition," *Proc. INTERSPEECH*, pp. 1094–1097, August 2007.

[17] R. Bold and A. MacDonald, "Theory of speech masking by reverberation," *Journal of the Acoustical Society of America*, vol. 21, no. 6, pp. 577–580, 1949.

[18] H. Kuttruff, *Room Acoustics*, 4th ed. London, UK: Spon Press, 2000.

[19] "HTK webpage," http://htk.eng.cam.ac.uk/.

[20] A. Sehr, R. Maas, and W. Kellermann, "Reverberation model-based decoding in the logmelspec domain for robust distant-talking speech recognition," *accepted for IEEE Transactions on Audio, Speech, and Language Processing Processing*.

[21] "Sound scene database in real acoustical environments," Real World Computing Partnership, 2001.

[22] R. Leonard, "A database for speaker-independent digit recognition," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 42.11.1–42.11.4, 1984.