Proceedings ESSV 2003

IMPORTANCE OF EARLY AND LATE REFLECTIONS FOR AUTOMATIC SPEECH RECOGNITION IN REVERBERANT ENVIRONMENTS

Heiko Gölzer and Michael Kleinschmidt

Medizinische Physik, Carl von Ossietzky Universität Oldenburg, D-26111 Oldenburg
Heiko.Goelzer@mail.uni-oldenburg.de; Michael.Kleinschmidt@uni-oldenburg.de
http://medi.uni-oldenburg.de/projects/asr

Abstract: In the European norm DIN EN ISO 3382 [1] about the "measurement of reverberation time of rooms with hints to other acoustical parameters" an early-to-late energy ratio is defined as a parameter that represents a ratio of early reflections energy to energies of reflections arriving after a certain critical delay time. The background understanding consists of the idea that in reverberant room situations early reflections improve the auditory perception while late portions, arriving with more than the critical delay time, have a detrimental effect. According to [1], the usual critical delay times for speech and music perception are 50 ms and 80 ms, respectively. This paper evaluates the importance of early and late reflections for the accuracy of automatic speech recognition (ASR). The effective time cutoff between conducive and detrimental portions of the impulse response is determined. This is done by successively deleting certain components of a room impulse response by setting them to zero. These modified impulse responses serve for reverberation of speech material that is recognized with a recognizer previously trained on non-reverberated speech. Canceling out portions with a detrimental effect for the recognition should cause increasing recognition accuracies, while eliminating conducive portions should result in lower accuracy rates. In this sense, recognition performance is observed dependent on manipulations of the underlying room impulse response. The results are substantially comparable to human perception: Early reflections up to a certain critical delay time can carry useful information and contribute to the recognition accuracy when detrimental late portions are present. This cutoff time is found in the range of 25 ms to 50 ms for different room impulse responses and two different front ends.

1 Introduction

In reverberant room environments a speech signal emitted from one location reaches the receiver on several paths first directly and then through reflections from cavity boundaries and objects like furniture. In general this reverberant distortion is assumed to have a detrimental effect on both human speech intelligibility and automatic speech recognition (ASR) accuracy.

At the same time effects are known that suggest a somewhat more detailed perspective on the processing of reverberant speech for human listeners. A widely known phenomenon in human speech perception is the temporal integration of the auditory system in reverberant room situations. Petzold [2] found a value of

$$t = 50ms \pm 10ms$$

for the so called “threshold of masking”. The value of delay time, above which a noticeable alteration of the acoustical impression occurs. Early reflections arriving within the first 50 ms after the direct sound are not perceived separately but are rather integrated for directional cues. A measure to characterize a reverberant room situation with respect to speech intelligibility based on this concept is defined in DIN EN ISO 3382 [1]. It is the early-to-late energy ratio

$$C_{t_e} = 10 \log_{10} \left(\int_0^{t_e} p^2(t)\, dt \;\Big/\; \int_{t_e}^{\infty} p^2(t)\, dt \right) [dB] \tag{1}$$

where $p^2(t)$ is the squared room impulse response and t_e is a characteristic critical delay time in milliseconds. C_{t_e} is a measure of the energy ratio between early and late reflections and therefore low for highly reverberant rooms and vice versa. According to [1], the usual measures for speech and music perception are C_{50} and C_{80} with time cutoffs of 50 ms and 80 ms, respectively. For human listeners the optimal cutoff might depend on the reverberant room situation and is, as the above separation implies, dependent on the signal itself.

A previous study was aimed at characterizing room situations [3]. While searching appropriate parameters to be estimated, the question arose which time cutoff would be optimal for automatic speech recognition. In this paper several experiments are carried out, searching for characteristic critical delay time for two different front ends. The experimental setup is to manipulate room impulse responses by setting individual portions to zero. These altered impulse responses are then used for convolving speech material to be recognized by ASR systems that are previously trained on non-reverberant conditions.

2 Experimental paradigm

The procedure used here follows the same path for four different experimental setups which are distinguished by different manipulations of underlying room impulse responses. In each case portions of the impulse response are successively set to zero to estimate their importance for the recognition task. The altered impulse responses are used for convolving a speech database by means of artificial reverberation. An automatic speech recognizer that was previously trained on non-reverberated speech is used for recognition in each artificial condition. Canceling out portions with a detrimental effect for the recognition should cause increasing recognition accuracies, while eliminating conducive portions should result in lower accuracy rates. In this sense the presented research considers the recognition accuracy of the system in relation to manipulations of the underlying room impulse response.

2.1 Manipulation of impulse responses

The importance of particular reflections for speech recognition results is always relative to other reflections present in the impulse response at earlier or later time instants. Therefore four different methods are chosen to investigate the importance of certain portions of the impulse response in different contexts (c.f. Fig. 1).

1. In the task named “growing gap” the reflections are successively set to zero starting at 5 ms after the direct maximum of the impulse response, up to a variable time instant t. This results in a gap of variable length in the impulse response. The portion of interest appears with leading zeros and in context with following reflections.

2. For the task named “moving gap” a 5 ms gap is introduced to the impulse response at different points in time starting 5 ms after the direct maximum and up to a variable time t.

This results in a gap in the impulse response at variable time instances with context of preceding and following reflections.

3. In the simplest task named “cutting tail” portions of the impulse response are removed from a variable time instant to the end. The resulting impulse response consists of a direct path and the following reflections up to the cutoff. The portion of interest is found in context with its preceding reflections.

4. The last task named “filling gap” is based on a gap of 95 ms starting 5 ms after the direct maximum. Portions of the impulse response are successively placed back into their original position, filling the gap from beginning to end. The portion of interest appears in context with the preceding reflections like for the “cutting tail” task but with later reflections being present.

Each manipulation task consists of a sequence of conditions with a shift of 5 ms from one condition to the next. The beginning and end of each gap is introduced by a Hanning window flank of 2 ms to ensure smooth transitions. The reference zero point for the time index is always the maximum value sample in the direct path of the impulse response. The different manipulations are depicted for a time index of 30 ms in Figure 1. The alignment is chosen to allow a comparison of the four different tasks. The time index gives the beginning time of the 5 ms portion of interest for a particular condition. For “growing gap” and “moving gap” it is the portion that has been removed from the precedent condition to the current one. For the tasks “cutting tail” and “filling gap” it is the portion that was introduced from the precedent condition to the current one.

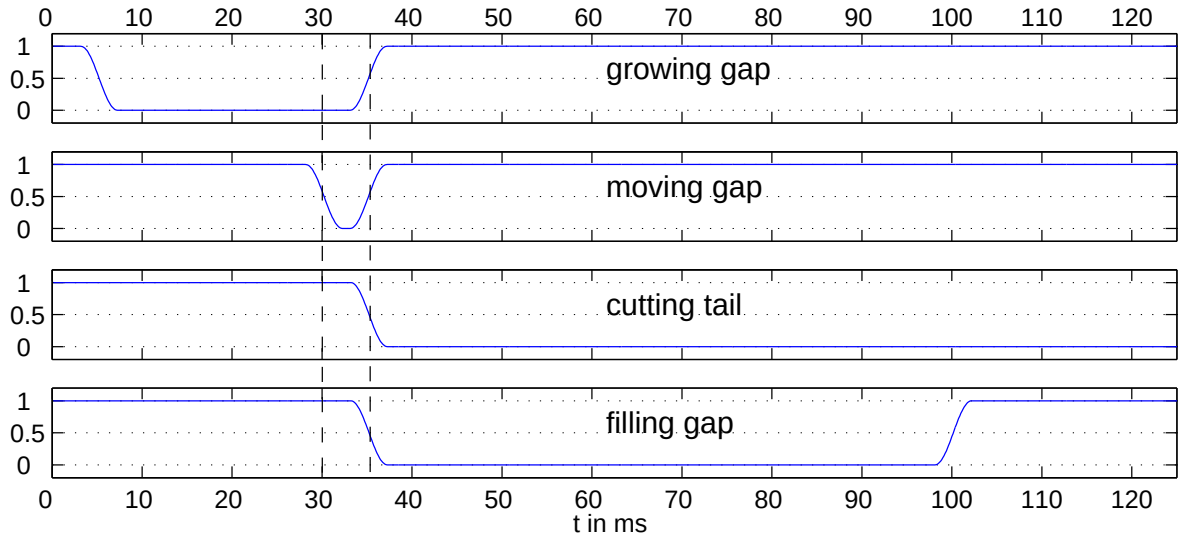


Figure 1 - *Manipulation of impulse responses by setting portions successively to zero. The gain factor for a time index of 30 ms is displayed for the four different tasks: a) growing gap, b) moving gap, c) cutting tail, d) filling gap. The time index refers to the starting time of the relevant portion. See Section 2.1 for details.*

2.2 Database of room impulse responses

For this experiment, a set of different room impulse responses is used. Four of them have been recorded in real reverberant enclosures and utilized by Couvreur et al. [4]. Their measurements

were based on a correlation method with optimal time-stretched pulses. One room impulse response was measured in each of four enclosures: cafeteria, lavatory, meeting room and office room. Another impulse response measured with maximum length sequences via a dummy head in an empty office room in Oldenburg is also part of the database (office OL). Additionally a matlab toolbox from DSP Algorithms [5] based on a mirror sound source model [6] is used to simulate room impulse responses in order to provide a broader base of room conditions. Two of these are represented here along with the measured room impulse responses. Reverberation time T_{60} and early-to-late energy ratio C_{30} of all room impulse responses are given in Table 1.

	impulse response						
	office	meeting	lavatory	cafeteria	office OL	sim(1)	sim(2)
T_{60}/s	0.74	1.10	1.72	1.88	0.78	1.35	1.40
C_{30}/dB	4.6	3.1	-4.0	2.6	4.4	3.4	0.3

Table 1 - *Reverberation time T_{60} and early-to-late energy ratio C_{30} of all room impulse responses used in this research.*

2.3 Speech material and recognizer setup

Speech material

The speech material for this experiment is part of the TIDIGITS database [7] as found in the AURORA2 framework [8]. It consists of digit sequences uttered by male and female native speakers of American English. The sound-files are sampled at 8 kHz with 16 bit quantization. Only the clean parts labeled as “train” and “testa” are used here. The database consists of 8440 and 4004 sequences for training and testing, respectively, with one to seven digits per sequence.

Recognizer setup

The system used for automatic speech recognition is a Hidden Markov based recognizer (HTK [9]) that is known from the AURORA2 contest [8]. The features in use are a) mel-frequency cepstral coefficients (MFCC), which are the reference processing in Aurora2 and b) log-RASTA-PLP [10]. For both front ends 13 cepstral features with dynamic Δ and $\Delta\Delta$ features [11] are used. The training is once carried out for both front ends on non-reverberated data using the train schedule proposed for the AURORA2 contest.

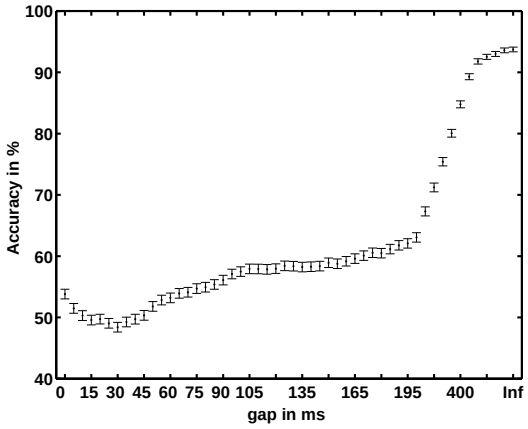
3 Results

The results are displayed as recognition accuracies in percent over a time index that represents the beginning of one portion in the room impulse response whose importance for the recognition task is considered in different contexts of earlier or later reflections. The error-bars indicate the standard deviation of a binomial distribution with $N = 4004$, the number of examples, and $p = \frac{Acc/\%}{100}$, the normalized recognition accuracy.

3.1 Front End Comparison

The two different front ends (MFCC and log-RASTA-PLP) show essentially no qualitative different behaviour concerning the characteristic temporal structure of detrimental and beneficial portions. As an example Figure 2 displays a comparison of the two different front ends with

a) office OL, MFCC



b) office OL, RASTA-PLP

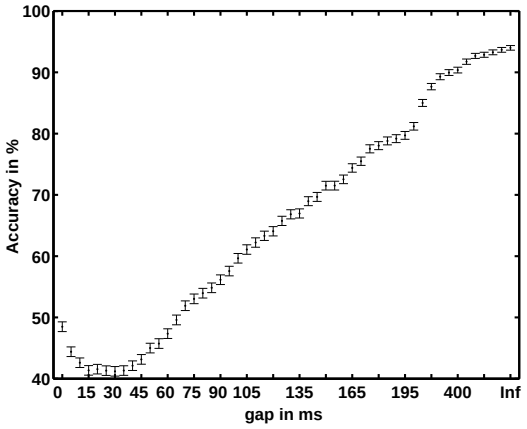


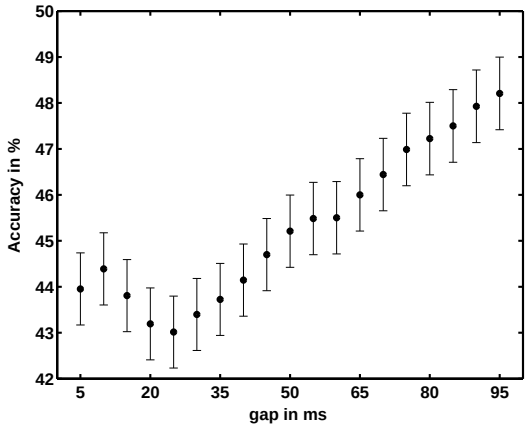
Figure 2 - *Recognition accuracies in percent for the "growing gap" task for two different front ends: a) MFCC. b) log-RASTA-PLP. The time index gives the length of the gap in milliseconds. At time index zero, the complete impulse response is used for reverberation. For the last condition labeled "Inf", only the direct path was used. The step size from one condition to the next is 5 ms up to 200 ms and then 40 ms up to the end of the impulse response.*

the "growing gap" task. Additionally the accuracies for the whole impulse response ("0") and for the direct path only ("Inf") are given. The expected behaviour for successively removing reflections from the room impulse response, an increase of recognition accuracy, can be found for later reflections on the right hand side in both figures. The interesting result for earlier time instances is a decrease of accuracy on removal of portions, which indicates beneficial reflections in the room impulse response. These appear up to a time of around 30 ms for the given context and the underlying impulse response. It is interesting to note, that the critical delay time is independent of front end type, although especially RASTA employs a long-term temporal filtering. This is also true for different types of delta-processing (not shown). In addition, the minimum is far above the half-window-length (12.5 s), ruling out any zero order effect. It can therefore be assumed that the results reflect some basic proporties of speech and/or the classification algorithm.

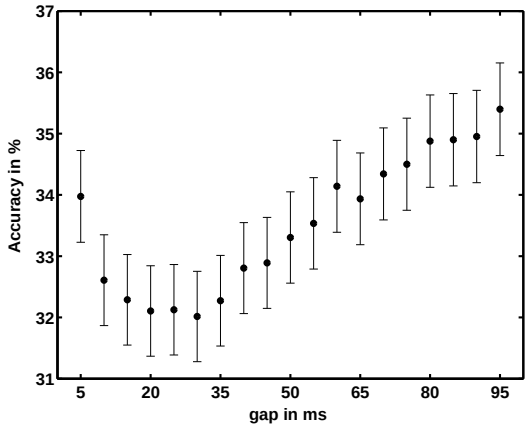
3.2 Room dependency

A comparison of different impulse responses under the "growing gap" task with MFCC features is presented in Figure 3. The effect of conducive early reflections is found for most room impulse responses. The time cutoff between beneficial and detrimental portions appears to be the highest for a room condition with very low early-to-late energy ratio (lavatory). In the highly reverberant cafeteria situation, early reflections don't seem to enhance the recognition performance. For somewhat more "ordinary" room conditions (meeting room, office room and office OL) with reverberation times roughly around 1 s and moderate early-to-late energy ratio, the minimum is located between 25 and 30 ms. For simulated rooms the effect is more pronounced, probably due to a more controlled situation with lacking noise floor in the impulse responses.

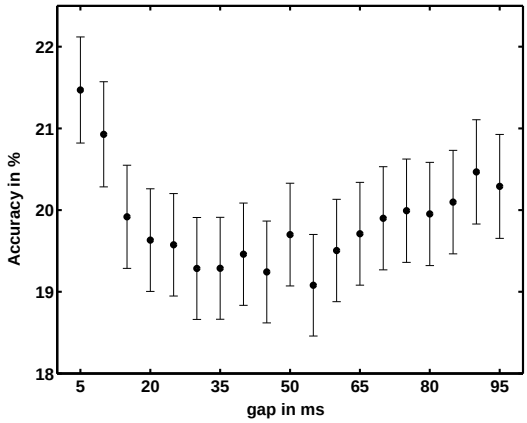
a) office room, MFCC



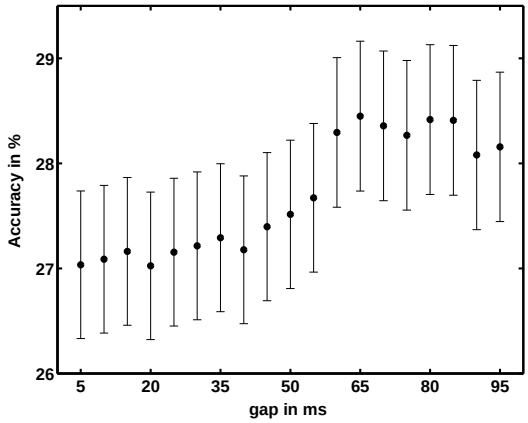
b) meeting room, MFCC



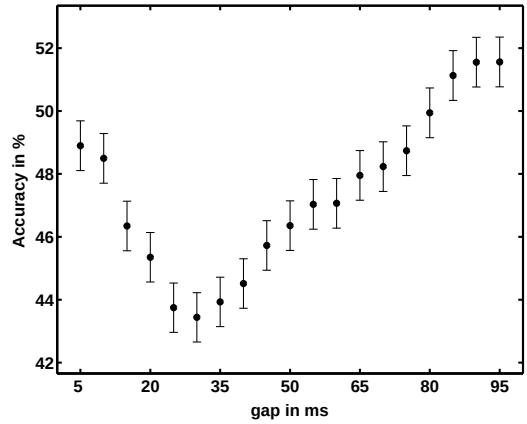
c) lavatory, MFCC



d) cafeteria, MFCC



e) simulated Room(1), MFCC



f) simulated Room(2), MFCC

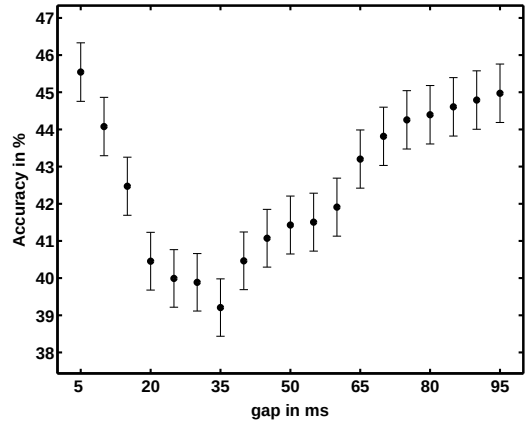


Figure 3 - *Recognition accuracies in percent for the "growing gap" task with six different room impulse responses. a) – d) measured room impulse responses; e) and f) simulated room impulse responses. The time index gives the length of the gap in milliseconds.*

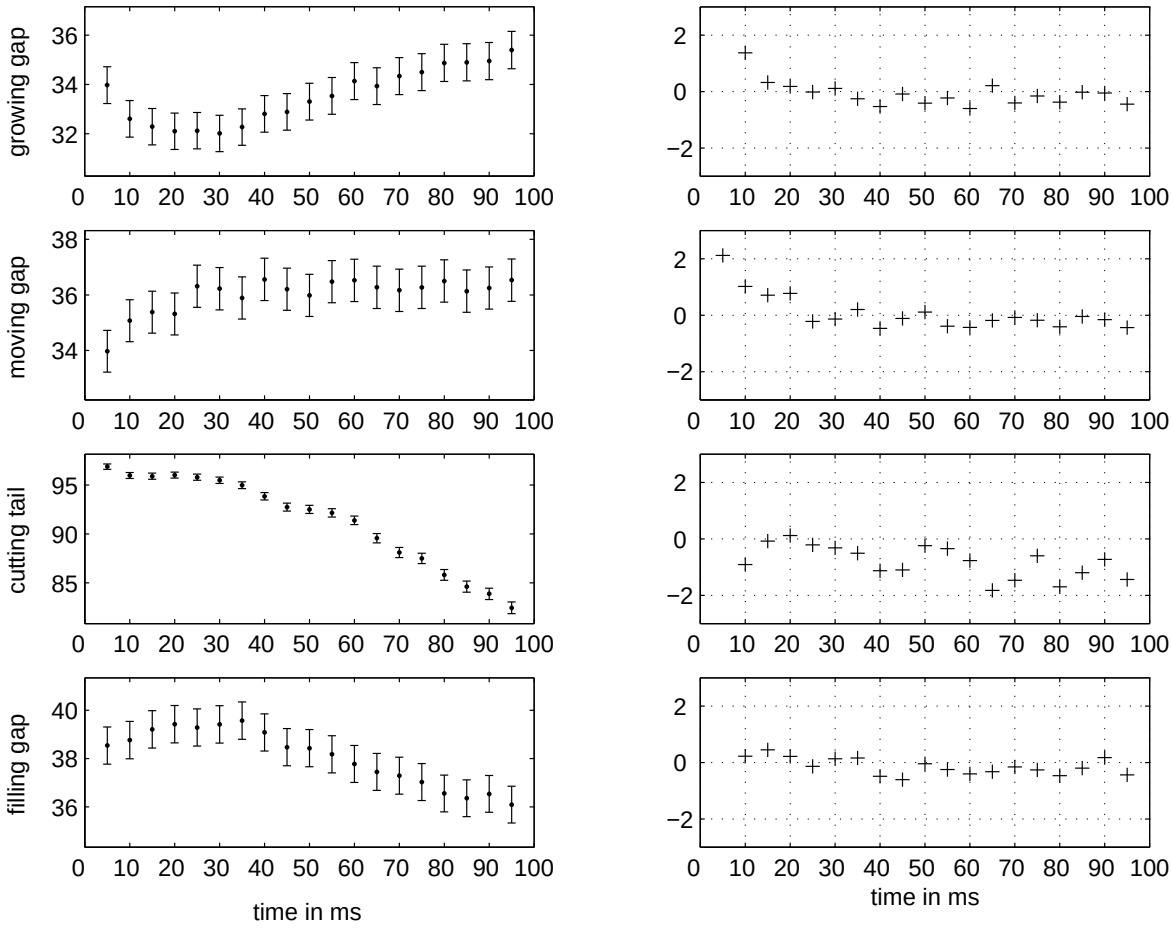


Figure 4 - *Results for the four different manipulation tasks. The left hand side displays recognition accuracies in percent for each task over a time index that represents the beginning time of the relevant 5 ms portion. The right hand side shows the difference of recognition accuracies from one time index to the next for the tasks "growing gap", "cutting tail" and "filling gap". A deviating illustration is chosen for the "moving gap" task, where accuracies are compared to the accuracy value under influence of the complete impulse response. A positive value is found for portions with a beneficial effect for the accuracy within their respective context. See Figure 1 and Section 2.1 for a better understanding of the underlying manipulations.*

3.3 Paradigm Evaluation

In Figure 4 results are displayed for all four tasks. The left side shows recognition accuracies in percent over a time index that refers to the beginning time of the relevant 5 ms portion in the impulse response. The right hand side displays the difference of recognition accuracy between one and the previous condition for the three tasks "growing gap", "cutting tail" and "filling gap". The accuracies for the "moving gap" task are compared to the recognition accuracy for testing with speech material reverberated with the complete impulse response. It is important to note that reading the figures from left to right, the upper task represent a successive removal of reflections with an expected increase of recognition accuracy, at least in the long run. For the two lower tasks, the opposite is the case. The difference plot on the right gives an overview of the importance of certain portions of the impulse response for speech recognition accuracy. A positive value is found for reflections in the impulse response that have a beneficial effect in their respective context and vice versa. Overall, the four experimental paradigms yield consistent

results. The 'growing gap' and the 'moving gap' paradigm give the clearest picture concerning the question of the correct choice for the cut-off time in ASR.

4 Summary

It was shown that the effect found for human auditory perception concerning the relevance of early and late reflections for speech perception in reverberant room situations, is under certain conditions also present for automatic speech recognition. Up to a critical delay time early reflections have a conducive effect on recognition accuracy, when late reflections are strongly represented in the room impulse response. When the effect is present, the characteristic time cutoff depends on the manipulation task and the underlying room impulse response and lies in a range between 25 ms and 50 ms. Both feature types (MFCC and RASTA-PLP) show the same qualitative behavior. In addition, the presented method allows for a detailed analysis of the influence of reverberation on recognition accuracy.

We thank Laurent Couvreur and Jörn Otten for providing us with room impulse responses. Special thanks go to Rainer Beutelmann, Bernd Meyer and Birger Kollmeier for their support.

Literature

[1] DIN EN ISO 3382, "Messung der Nachhallzeit von Räumen mit Hinweisen auf andere raumakustische Parameter," 2000.

[2] Petzold, Elementare Raumakustik, *Berlin, 1927.*

[3] H. Gölzer and M. Kleinschmidt, "Automatische Schätzung wichtiger Nachhallparameter," in Fortschritte der Akustik - DAGA, DEGA, *2003.*

[4] L. Couvreur, C. Ris, and C. Couvreur, "Model-based Blind Estimation of Reverberation Time: Application to Robust ASR in Reverberant Environments," in Proc. Eurospeech, *Aalborg, Denmark, 2001, vol. 4, pp. 2631–2634.*

[5] DSP Algorithms, "Room v.2, matlab scripts," 2000.

[6] J.B. Allen and D.A. Berkley, "Image method for efficiently simulating small-room acoustics," JASA, *vol. 65, no. 4, pp. 943–950, 1979.*

[7] R.G. Leonard and G. Doddington, "Tidigits speech corpus," Texas Instruments, Inc, 1993.

[8] H.G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noise conditions," ISCA ITRW ASR2000, 2000.

[9] HTK, "Hidden markov model toolkit v3.1," http://htk.eng.cam.ac.uk/, 2002.

[10] H. Hermansky and N. Morgan, "RASTA processing of speech," IEEE Trans. SAP, *vol. 2, no. 4, pp. 578–589, 1994.*

[11] S. Furui, "Speaker independend isolated word recognition using dynamic features of speech spectrum," IEEE Trans. Acoust., Speech, Signal Processing, *vol. 34, pp. 52–59, 1986.*