MAXIMUM LIKELIHOOD ESTIMATION OF A REVERBERATION MODEL FOR ROBUST DISTANT-TALKING SPEECH RECOGNITION

Armin Sehr[1], *Yuanhang Zheng*[1], *Elmar Nöth*[2] *and Walter Kellermann*[1]

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

[2] Pattern Recognition (Informatik 5),
University of Erlangen-Nuremberg
Martensstr. 3, 91058 Erlangen, Germany
noeth@informatik.uni-erlangen.de

ABSTRACT

We propose a novel approach for estimating a reverberation model for a robust recognizer according to [1], which is designed to allow distant-talking automatic speech recognition (ASR) in reverberant environments. Based on a few calibration utterances with known transcriptions recorded in the target environment, a maximum likelihood estimator is used to find the means and variances of the reverberation model. In contrast to [1] and to HMM training on artificially reverberated training data (e. g. [2]), measurements of room impulse responses become unnecessary, and the effort for training is greatly reduced. Simulations of a connected digit recognition task show that, in highly reverberant environments, the reverberation models estimated by the proposed approach achieve significantly higher recognition rates than HMMs trained on reverberant data.

1. INTRODUCTION

Current state-of-the-art ASR systems work reliably only if close-talking microphones are used for the speech input. This is a major acceptance problem of current ASR applications because most users find it uncomfortable to wear a headset or to use any other close-talking microphone. Therefore, reliable distant-talking ASR is highly desirable.

Since the distance between speaker and microphone in a distant-talking scenario usually is in the range of one to several meters, the microphone does not only capture the desired signal, but also unwanted additive signals and reverberation of the desired signal, both of which hamper ASR. While significant progress has been achieved in the last decades in improving the robustness of ASR to additive distortions, the research on reverberation-robust ASR is still in its infancy. This paper focuses on robustness to reverberation.

The reverberant speech signal $x(t)$ is given by the convolution of the clean speech signal $s(t)$ with the room impulse response RIR $h(t)$ describing the acoustic path between speaker and microphone

$$x(t) = h(t) * s(t)\ .$$

For typical reverberant environments like offices or living rooms, the length of the RIR is in the range of 300 ms to 800 ms. Thus the RIR is much longer than the speech frames used for feature extraction with a typical length of about 20 ms and a typical frame shift of about 10 ms. That is, the RIR extends over a large number of speech frames. Therefore, the effect of reverberation cannot be modeled as a simple multiplication or addition in the feature domain.

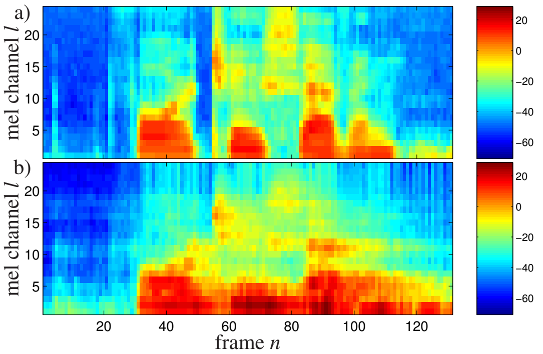

Figure 1: a) Clean and b) reverberant mel-spectral feature vector sequences corresponding to the utterance ”four, two, seven” using a dB color scale.

Figure 1, comparing the feature sequences of the clean (recorded by a close-talking microphone) and the reverberant (microphone four meters away from the speaker) utterance ”four, two, seven” in the mel-spectral domain, illustrates that reverberation has a dispersive effect on the speech feature sequences: The features are smeared along the time axis so that the current feature vector depends strongly on the previous feature vectors. We believe that this contradiction to the conditional independence assumption of HMMs ([3], chapter 8), namely that the current feature vector depends only on the current state, implies a major performance limitation of HMM-based recognizers in reverberant environments.

The dispersive effect of reverberation is the reason why conventional model adaptation approaches developed mainly for the adaptation to additive distortions (see e. g. [3], chapter 10 for an overview), cepstral mean subtraction [4] or the use of Δ and $\Delta\Delta$ features [5] are not very effective in reverberant environments. All these approaches effectively improve recognition performance when the speech signal is convolved by a short impulse response. However they achieve only limited improvements with the long impulse responses of room reverberation.

To improve the robustness of HMM-based recognizers to reverberation, model training with artificially reverberated training data [2, 6, 7] and model adaptation approaches tailored particularly to reverberation [8, 9] have been proposed. Both methods achieve a significant increase in recognition rate in reverberant environments compared to HMMs trained only on clean data. However, since both methods still solely rely on HMMs, their performance is limited by the conditional independence assumption. Furthermore, the reverberant training approach implies a considerable effort. Room impulse responses, used to generate the reverberant training

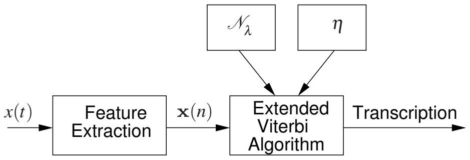

Figure 2: ASR based on a combination of an HMM network $\mathcal{N}_\lambda$ and a reverberation model η according to [1].

data, have to be measured in the target environment, and a complete training has to be performed.

Recently, two approaches have been proposed to overcome the limitation of the conditional independence assumption. In [10] a frame-by-frame adaptation method is suggested which estimates the reflection of the previous feature vectors by a first-order linear prediction and adds the estimate to the means of the clean-speech HMM. This implies an approximation of the reverberation by a strictly exponentially decaying function and achieves slightly lower recognition rates compared to matched reverberant training [10].

The approach proposed in [1] uses a combination of an HMM network and a reverberation model to describe the reverberant feature sequence. In this way, a very accurate model of the reverberation is achieved so that the approach outperforms conventional HMM-based recognizers trained on matched reverberant speech. However, the training of the reverberation model suggested in [1] still requires the measurement of room impulse responses (RIRs) in the room where the recognizer is to be used.

In some important applications, measuring a set of room impulse responses in the target environment is either not possible or too expensive. Therefore, in this paper, we propose a new way to estimate the reverberation model directly in the feature domain. The feature-domain representation of the RIRs in the target environment is determined by maximum likelihood (ML) estimation based on the reverberant feature sequences of a few calibration utterances with known transcriptions. In this way, measurements of room impulse responses become unnecessary, and the effort for training is greatly reduced compared to [1] and compared to the reverberant training approaches.

The paper is organized as follows: In Section 2, the approach proposed in [1] is reviewed with a special emphasis on the reverberation model. The new estimation of the reverberation model is derived in Section 3 and simulation results are discussed in Section 4. In Section 5, the paper is summarized and conclusions are drawn.

2. REVERBERATION MODEL FOR ROBUST ASR

To model the reverberant feature sequences without the limitation imposed by the conditional independence assumption, [1] suggests to use a combination of an HMM network $\mathcal{N}_\lambda$ modeling the clean speech and a reverberation model η as depicted in Figure 2. The combination of the model outputs is performed directly in the feature domain.

The combination operator for the two model outputs is feature-dependent. A very simple combination operator can be used for mel-spectral (melspec) coefficients, which are a preliminary stage in the calculation of the MFCC coefficients without the logarithm and the DCT operation as depicted in Figure 3. Therefore, melspec coefficients are used as speech features throughout this paper, even though they cannot be modeled very well by single Gaussian densities.

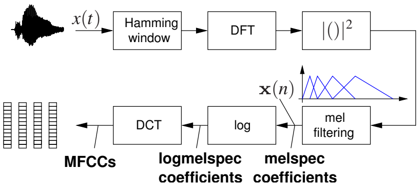

Figure 3: Calculation of melspec features.

In the melspec domain, the combination operation of the model outputs can be expressed as a convolution of the feature vectors

$$\mathbf{x}(n) = \sum_{m=0}^{M-1} \mathbf{h}(m,n) \odot \mathbf{s}(n-m) \quad \forall\, n = 1 \ldots N+M-1\,. \tag{1}$$

where $\mathbf{x}(n)$, $\mathbf{h}(m,n)$ and $\mathbf{s}(n-m)$ are the reverberant feature vector, the output of the reverberation model and the output of the clean-speech HMM network, respectively.

The reverberation model can be thought of as a feature domain representation of the RIR. However, the reverberation model does not only represent a single fixed RIR. It rather is a statistical representation of all possible RIRs of the room where the recognizer will be used.

The reverberation model exhibits a matrix structure where each row corresponds to a certain mel channel and each column to a certain frame as shown in Figure 4. The matrix elements are modeled by random variables. For simplicity, these random variables are assumed to be statistically independent and normally distributed.

At each time frame, each random variable produces a new realization which is statistically independent from the previous realizations. In this way, the reverberation model can be considered as a matrix-valued independent identically distributed (IID) random process.

For the recognition, an extended version of the Viterbi algorithm is employed to find the most likely path through the network of HMMs in connection with the reverberation model. At each Viterbi iteration

$$\gamma_j(n) = \max_i \{\gamma_i(n-1) \cdot a_{ij} \cdot O(n)\}\,,$$

an inner optimization

$$O(n) = \max_{\mathbf{s}(n),\mathbf{h}(m,n)} \{\, f_\lambda(j,\mathbf{s}(n)) \cdot f_\eta(\mathbf{h}(0,n),\ldots,\mathbf{h}(M-1,n))\} \tag{2}$$

has to be performed in order to find the optimum combination of the HMM output and the reverberation model output. Here, $\gamma_j(n)$, a_{ij} and $O(n)$ denote the Viterbi score of frame n and state j, the transition probability from state i to state j and the output density of the combined model, which is given by maximizing the product of the HMM output density $f_\lambda(j,\mathbf{s}(n))$ and the reverberation model output density $f_\eta(\mathbf{h}(0,n),\ldots,\mathbf{h}(M-1,n))$. Further details are given in [1].

To train the reverberation model, [1] proposes to use a set of room impulse responses measured for different loudspeaker and microphone positions in the room where the recognizer will be used. After time alignment, the feature domain representations of all RIRs are calculated and used to determine the means and variances of the reverberation model.

3. MAXIMUM LIKELIHOOD ESTIMATION OF THE REVERBERATION MODEL

In this section, a novel approach for estimating the reverberation model which does not require the measurement of room

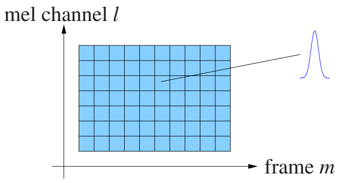

Figure 4: Reverberation model.

impulse responses is derived. The reverberation model is estimated directly in the feature domain based on a few calibration utterances with known transcriptions recorded in the target environment by the recognizer's distant microphone.

For each utterance, the reverberant speech signal is transformed to the feature domain yielding the reverberant feature sequence $\mathbf{x}(1)\dots\mathbf{x}(N)$, where

$$\mathbf{x}(n) = [x_1(n), x_2(n), \dots, x_L(n)]^T$$

is the $L \times 1$ reverberant feature vector at frame n, and $x_l(n)$ is the l-th feature of $\mathbf{x}(n)$, N is the length of the utterance and L is the number of features per vector.

The reverberation model is estimated as follows. The reverberant feature sequence $\mathbf{x}(1)\dots\mathbf{x}(N)$ is segmented into hyper-frames with a length of K and a relative shift of P frames. For each hyper-frame, a speech model describing the clean-speech hyper-frame is determined. From this clean speech model, a reverberant speech model is derived.

The idea now is to find an estimate of the melspec RIR representation for the current hyper-frame by maximizing the probability of the reverberant hyper-frame given the reverberant speech model. Using the melspec RIR representations for all hyper-frames, the means and variances of the reverberation model are calculated as described below.

For simplicity, the speech features are assumed to be statistically independent. Therefore, the estimation of the reverberation model can be performed separately for each mel channel. We define the $1 \times K$ vectors

$$\begin{aligned}\mathbf{s}_l(k) &= [s_l(kP), s_l(kP+1), \dots, s_l(kP+K-1)]\\ \mathbf{x}_l(k) &= [x_l(kP), x_l(kP+1), \dots, x_l(kP+K-1)]\end{aligned}$$

representing the clean and reverberant features of the l-th mel channel of hyper-frame k.

Using the HMMs of the recognizer, which are trained on clean speech, and the known transcriptions, a sequence of HMMs describing the word sequence of the current calibration utterance is constructed. The most likely path through this HMM sequence is determined by the Viterbi algorithm. In this way, the state/frame-alignment between the reverberant feature sequence and the HMM sequence is obtained. That is, a certain HMM state is assigned to each frame n of the reverberant sequence. Using the output densities of the aligned HMM states, the joint probability density

$$f_{\mathbf{S}_l(k)}(\mathbf{s}_l(k)) = f_{S_l(kP)}(s_l(kP)) \cdot \dots \cdot f_{S_l(kP+K-1)}(s_l(kP+K-1))$$

for the l-th feature of the clean-speech hyper-frame k is obtained, where $\mathbf{S}_l(k)$ and $S_l(kP+n)$ are the random vector and random variable describing $\mathbf{s}_l(k)$ and $s_l(kP+n)$, respectively. Note that statistical independence of all frames is assumed for simplicity.

The joint density for the l-th feature of the reverberant hyper-frame k is approximated by

$$\begin{aligned}f_{\mathbf{X}_l(k)}(\mathbf{x}_l(k)) &= f_{X_l(kP)}(x_l(kP)) \cdot \dots \cdot f_{X_l(kP+K-1)}(x_l(kP+K-1))\\ &= A \cdot \exp\left(-\frac{1}{2}(\mathbf{x}_l(k) - \mu_{\mathbf{X}_l(k)})\, C^{-1}_{\mathbf{X}_l(k)\mathbf{X}_l(k)}(\mathbf{x}_l(k) - \mu_{\mathbf{X}_l(k)})^T\right),\end{aligned}$$

where A is a normalizing constant.

The $1 \times K$ mean vector $\mu_{\mathbf{X}_l(k)}$ is approximated by a convolution in the feature domain as

$$\begin{aligned}\mu_{\mathbf{X}_l(k)} &= [\mu_{X_l}(kP), \dots, \mu_{X_l}(kP+K-1)]\\ &= \mathbf{h}_l(k)\, \mu_{\mathbf{S}_l(k)},\end{aligned}$$

where

$$\mu_{\mathbf{S}_l(k)} = \begin{bmatrix} \mu_{S_l}(kP) & \mu_{S_l}(kP+1) & \cdots & \mu_{S_l}(kP+K-1)\\ 0 & \mu_{S_l}(kP) & \cdots & \mu_{S_l}(kP+K-2)\\ 0 & 0 & \cdots & \mu_{S_l}(kP+K-3)\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \cdots & \mu_{S_l}(kP+K-M) \end{bmatrix}$$

is a $M \times K$ matrix containing the means of feature l from the clean speech model needed to calculate the corresponding means of the reverberant model for hyper-frame k, and

$$\mathbf{h}_l(k) = [h_l(k,0), \dots, h_l(k,M-1)]$$

is the $1 \times L$ vector describing the l-th mel channel of the feature domain RIR representation at hyper-frame k. Here, the m-th vector element $h_l(k,m)$ represents the m-th frame of this RIR representation.

For simplicity, the $K \times K$ diagonal covariance matrix $C_{\mathbf{X}_l(k)\mathbf{X}_l(k)}$ of the reverberant hyper-frame $\mathbf{X}_l(k)$ is approximated by the covariance matrix $C_{\mathbf{S}_l(k)\mathbf{S}_l(k)}$ of the clean hyper-frame $\mathbf{S}_l(k)$ as suggested in [8]

$$C_{\mathbf{X}_l(k)\mathbf{X}_l(k)} = C_{\mathbf{S}_l(k)\mathbf{S}_l(k)}.$$

To find the most likely RIR representation for hyper-frame k, $f_{\mathbf{X}_l(k)}(\mathbf{x}_l(k))$ is maximized with respect to $\mathbf{h}_l(k)$ for each mel channel $l = 1\dots L$. Taking the logarithm of $f_{\mathbf{X}_l(k)}(\mathbf{x}_l(k))$ and neglecting the irrelevant constants A and $-1/2$, we obtain the cost function

$$J(\mathbf{h}_l(k)) = (\mathbf{x}_l(k) - \mu_{\mathbf{X}_l(k)})\, C^{-1}_{\mathbf{X}_l(k)\mathbf{X}_l(k)}(\mathbf{x}_l(k) - \mu_{\mathbf{X}_l(k)})^T.$$

Minimizing $J(\mathbf{h}_l(k))$ is equivalent to maximizing $f_{\mathbf{X}_l(k)}(\mathbf{x}_l(k))$, so the maximum likelihood estimate of $\mathbf{h}_l(k)$ is given by

$$\hat{\mathbf{h}}_l(k) = \underset{\mathbf{h}_l(k)}{\operatorname{argmin}}\ J(\mathbf{h}_l(k)).$$

Using $\mu_{\mathbf{X}_l(k)} = \mathbf{h}_l(k)\, \mu_{\mathbf{S}_l(k)}$ and $C_{\mathbf{X}_l(k)\mathbf{X}_l(k)} = C_{\mathbf{S}_l(k)\mathbf{S}_l(k)}$, the cost function can be expressed as

$$\begin{aligned}J(\mathbf{h}_l(k)) &= (\mathbf{x}_l(k) - \mathbf{h}_l(k)\, \mu_{\mathbf{S}_l(k)})\, C^{-1}_{\mathbf{S}_l(k)\mathbf{S}_l(k)}(\mathbf{x}_l(k) - \mathbf{h}_l(k)\, \mu_{\mathbf{S}_l(k)})^T\\ &= \mathbf{x}_l(k)\, C^{-1}_{\mathbf{S}_l(k)\mathbf{S}_l(k)}\, \mathbf{x}_l^T(k) - 2\, \mathbf{h}_l(k)\, \mu_{\mathbf{S}_l(k)}\, C^{-1}_{\mathbf{S}_l(k)\mathbf{S}_l(k)}\, \mathbf{x}_l^T(k)\\ &\quad + \mathbf{h}_l(k)\, \mu_{\mathbf{S}_l(k)}\, C^{-1}_{\mathbf{S}_l(k)\mathbf{S}_l(k)} \mu^T_{\mathbf{S}_l(k)}\, \mathbf{h}_l^T(k).\end{aligned}$$

	Room A	Room B	Room C
Type	lab	studio	lecture room
T_{60}	300 ms	700 ms	900 ms
d	2.0 m	4.1 m	4.0 m
SRR	4.0 dB	−4.0 dB	-4.0dB

Table 1: Summary of room characteristics: T_{60} is the reverberation time, d the distance between speaker and microphone and SRR is the signal-to-reverberation-ratio.

Calculating the derivative of the cost function with respect to $\mathbf{h}_l(k)$ and setting it to zero, we obtain

$$\hat{\mathbf{h}}_l^T(k) = (\mu_{\mathbf{S}_l(k)} C^{-1}_{\mathbf{S}_l(k)\mathbf{S}_l(k)} \mu^T_{\mathbf{S}_l(k)})^{-1} \mu_{\mathbf{S}_l(k)} C^{-1}_{\mathbf{S}_l(k)\mathbf{S}_l(k)} \mathbf{x}_l^T(k) . \quad (3)$$

The maximum likelihood melspec representations $\hat{\mathbf{h}}_l(k)$ are found in this way for all hyper-frames k and all mel channels l and are used to estimate the means and variances of the reverberation model

$$\mu_{\mathbf{h}_l} = \frac{1}{J} \sum_{k=1}^{J-1} \hat{\mathbf{h}}_l(k) \quad \text{for} \quad l = 1, \ldots, L, \quad (4)$$

$$\sigma^2_{\mathbf{h}_l} = \frac{1}{J-1} \sum_{k=1}^{J-1} \left(\hat{\mathbf{h}}_l(k) - \mu_{\mathbf{h}_l(k)} \right)^2 \text{ for } l = 1, \ldots, L, \quad (5)$$

where J is the number of hyper-frames. Thus, all parameters of the reverberation model are determined.

4. EXPERIMENTS

To verify how well the effect of reverberation can be described by the ML reverberation models determined according to Section 3, we perform experiments with reverberant versions of the TI digits corpus [11] (sampling rate 20 kHz) in three different rooms. The ML reverberation models are first compared to the reverberation models obtained from measured room impulse responses according to [1], which will be referred to as exact reverberation models in the following. Then, connected digit recognition (CDR) experiments are performed in all three rooms. The recognition performance of the ML reverberation models and the exact reverberation models used according to [1] are compared to those of conventional HMM-based recognizers.

4.1 ML Estimation of Reverberation Models

For the ML estimation of the reverberation models, calibration utterances with known transcriptions are used. Therefore, 20 utterances from the TI digits training set, containing 140 digits in total, are selected as calibration utterances and are convolved with the RIRs measured in three different rooms. The characteristics of these rooms are summarized in Table 1.

The reverberant calibration utterances are then transformed to the feature domain by calculating 24 melspec coefficients for each frame using a frame length of 25 ms, a frame shift of 10 ms and a DFT size of 512.

To get the clean speech model, two different sets of 16-state word-level HMMs are used, one set in the melspec domain, the other in the MFCC domain. Using the clean training set of the TI digits corpus, these HMM sets are trained in the following way. First, single Gaussian MFCC-based HMMs are trained by ten iterations of Baum-Welch re-estimation. Then, single Gaussian melspec HMMs are obtained from the MFCC-based HMMs by single pass retraining [12].

	Room A	Room B	Room C
K	length of the current calibration utterance		
J	20 utterances = 20 hyper-frames		
M	20	50	70

Table 2: Parameters used for the ML estimation of the reverberation model: K: hyper-frame length, J: number of hyper-frames, M length of the reverberation model in frames.

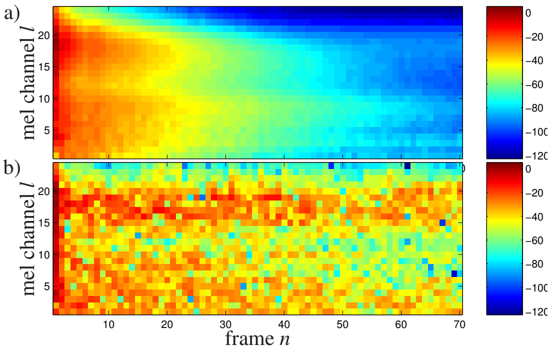

Figure 5: Mean values of a) the exact reverberation model b) the ML reverberation model for room C (melspec domain with dB color scale).

The MFCC-based HMMs are used to determine the state/frame-alignment between the feature vectors and the HMMs. Based on this state/frame-alignment, the joint probability density $f_{\mathbf{S}_l(k)}(\mathbf{s}_l(k))$ of the clean-speech hyper-frame k is obtained for all mel channels by assigning the corresponding single Gaussian output densities of the respective states of the melspec HMMs to the aligned frames.

Using the mean vector and the covariance matrix of this density, the means and variances of the reverberation model are calculated according to equations (3), (4) and (5). The parameter settings used for these calculations are summarized in Table 2. The hyper-frame length is set to the variable length of the utterances. That is, each of the 20 calibration utterances is used as one hyper-frame.

Figure 5 compares the mean values of the exact reverberation model to the mean values of the ML reverberation model for room C. The two images do not look similar at the first glance, since the details of their time/frequency-patterns are different. However, a closer inspection shows that the means obtained by ML estimation are a fairly good approximation to the basic envelope of the exact means for the first 20-30 frames. For example both time/frequency-patterns exhibit regions of high power around the mel channels 4 and 18 and regions of low power around channels 12 and 24.

For the later frames, the approximations of the ML approach lead to an overestimation of the reverberation. Therefore, in the second half of the reverberation model, the ML means are significantly higher than the exact means. Similar results are obtained for room A and B. Therefore, only the first half of the ML reverberation model is used in the following experiments.

4.2 Connected Digit Recognition Experiments

The recognition rates of the ML reverberation models are compared to that of the exact reverberation models and to that of conventional HMM-based recognizers by simulations of a connected digit recognition task.

	clean	Room		
	data	A	B	C
(I) conv. clean training	82.0	51.5	13.4	25.9
(II) conv. reverb. training	-	66.8	54.6	46.0
(III) exact reverberation model	-	77.6	71.6	67.6
(IV) ML reverberation model	-	63.0	57.3	58.5

Table 3: Word accuracies for the conventional HMM-based recognizer trained on clean (I) and reverberant speech (II) and for the reverberation model-based approaches according to [1] using exact reverberation models (III) and ML reverberation models (IV).

One set of test data is obtained for each room A, B and C by convolving the clean data from the TI digits test set with different RIRs measured in the corresponding room.

The clean-speech single Gaussian melspec HMMs as described in Section 4.1 are used both for the reverberation model-based approaches and the conventional HMM-based approach. Additionally, HMMs trained on matched reverberant data from the respective rooms are used for the conventional HMM-based approach.

As only the early frames of the ML estimation are reliable, only the first half of each ML reverberation model is used, so that the length of the ML reverberation model is $M = 10/25/35$ in room A/B/C. Also, the ML variance estimation according to equation (5) deviates significantly from the variances of the exact reverberation models. Therefore, the ML variance estimation is replaced by the square of the ML means, which has been found to represent a better approximation of the exact variance.

Table 3 compares the recognition rates of the HMM-based approaches using HMMs trained on clean (I) and matched reverberant data (II), and the reverberation model-based approaches using the exact (III) and the ML (IV) reverberation models.

With increasing reverberation, the word accuracy of the HMM-based recognizer trained on clean speech (I) decreases significantly. Even though, room B is less reverberant than room C, the performance of (I) is lower in room B, because room B exhibits a strong low-pass characteristic. Using HMMs trained on reverberant data of matched conditions (II), the recognition performance is increased considerably in all reverberant environments. Using exact reverberation models in the recognizer concept of [1] (III), the recognition rate is further increased, almost approaching the clean-speech performance of (I).

The word accuracy achieved with the ML reverberation models (IV) is significantly higher than that of (I). In the moderately reverberant room A, it is comparable to the accuracy achieved by the reverberant HMMs (II). In the strongly reverberant room C, the proposed ML reverberation models (IV) perform significantly better than (II). The reason why (IV) only slightly outperforms (II) in the highly reverberant room B is the low-pass characteristic of room B, which can be very well modeled by the conventional HMM-based recognizer (II).

In summary, using the ML reverberation models in the recognizer concept according to [1], a recognition performance comparable or better than that of reverberantly trained HMMs can be achieved without measuring room impulse responses in the target environment. Furthermore, these results confirm that the recognizer concept according to [1] is robust to inaccuracies in the reverberation model.

5. CONCLUSIONS

A novel approach for estimating the reverberation model, which is used in the recognizer concept according to [1] for robust distant-talking ASR in reverberant environments, has been proposed in this paper. A few calibration utterances with known transcriptions are recorded in the target environment and are used in a maximum likelihood estimation approach to find the means and variances of the reverberation model. The approach allows the reliable estimation of the early frames of the reverberation model without measuring room impulse responses in the target environment. In this way, the effort for training is greatly reduced compared to training HMMs on artificially reverberated data. Simulation results of a connected digit recognition task have shown that the proposed ML reverberation models achieve recognition rates comparable to those of reverberant HMMs in moderately reverberant rooms and significantly better than those of reverberant HMMs in strongly reverberant rooms.

REFERENCES

[1] A. Sehr, M. Zeller, and W. Kellermann, "Distant-talking continuous speech recognition based on a novel reverberation model in the feature domain," *Proc. International Conference on Spoken Language processing (ICSLP/INTERSPEECH)*, September 2006.

[2] D. Giuliani, M. Matassoni, M. Omologo, and P. Svaizer, "Training of HMM with filtered speech material for hands-free recognition," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 449–452, March 1999.

[3] X. Huang, A. Acero, and H.-W. Hon, *Spoken language processing: A guide to theory, algorithm, and system development*, Prentice Hall, Upper Saddle River, NJ, USA, 2001.

[4] 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.

[5] S. Furui, "On the role of spectral transition for speech perception," *Journal of the Acoustical Society of America*, vol. 80, no. 4, pp. 1016–1025, 1986.

[6] V. Stahl, A. Fischer, and R. Bippus, "Acoustic synthesis of training data for speech recognition in living-room environments," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 285–288, May 2001.

[7] T. Haderlein, E. Nöth, W. Herbordt, W. Kellermann, and H. Niemann, "Using Artificially Reverberated Training Data in Distant Talking ASR," in *Proc. 8th International Conference on Text, Speech and Dialogue (TSD 2005)*, V. Matoušek, P. Mautner, and T. Pavelka, Eds., Berlin, 2005, vol. 3658 of *Lecture Notes for Artificial Intelligence*, pp. 226–233, Springer–Verlag.

[8] C. K. Raut, T. Nishimoto, and S. Sagayama, "Model adaptation for long convolutional distortion by maximum likelihood based state filtering approach," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. I–1133 – I–1136, May 2006.

[9] H.-G. Hirsch and H. Finster, "A new HMM adaptation approach for the case of a hands-free speech input in reverberant rooms," *Proc. INTERSPEECH*, pp. 781–783, September 2006.

[10] T. Takiguchi, M. Nishimura, and Y. Ariki, "Acoustic model adaptation using first-order linear prediction for reverberant speech," *IEICE Transactions on Information and Systems*, vol. E89-D, no. 3, pp. 908–914, March 2006.

[11] R. G. 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.

[12] P. C. Woodland, M. J. F. Gales, and D. Pye, "Improving environmental robustness in large vocabulary speech recognition," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 65–68, May 1996.