A COMBINED APPROACH FOR ESTIMATING A FEATURE-DOMAIN REVERBERATION MODEL IN NON-DIFFUSE ENVIRONMENTS

Armin Sehr[1], Jimi Y. C. Wen[2], Walter Kellermann[1], and Patrick A. Naylor[2]

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

[2]Department of EEE
Imperial College, London, UK
{yung.wen,p.naylor}@imperial.ac.uk

ABSTRACT

A combined approach for estimating a feature-domain reverberation model suitable for the robust distant-talking automatic speech recognition concept REMOS (REverberation MOdeling for Speech recognition) [1] is proposed. Based on a few calibration utterances recorded in the target environment, the combined approach employs ML estimation and blind estimation of the reverberation time to determine a two-slope reverberation model. Since measurements of room impulse responses become unnecessary, the effort for training is greatly reduced compared to [1] and compared to training HMMs on artificially reverberated data. Connected digit recognition experiments show that the proposed reverberation models in connection with the REMOS concept significantly outperform HMM-based recognizers trained on reverberant data.

Index Terms— Dereverberation, blind estimation, reverberation model, reverberation time, robust ASR.

1. INTRODUCTION

Reverberation caused by multi-path propagation of sound waves from the source to the microphone in distant-talking scenarios does not only reduce the perceived sound quality but also decreases the performance of Automatic Speech Recognition (ASR) significantly [2]. To increase the robustness of ASR to reverberation, either the speech signal can be dereverberated before the features are extracted [3] or the acoustic models of the recognizer can be adapted to the reverberation [4].

The REMOS concept introduced in [1] combines both approaches by performing dereverberation directly in the feature domain based on an acoustic model consisting of clean-speech HMMs and a ReVerberation Model (RVM). Thus, an extremely robust recognizer is achieved outperforming conventional HMM-based recognizers trained on reverberant speech data matching the acoustic conditions of the application environment. However, the estimation method for determining the RVMs suggested in [1] requires the measurement of Room Impulse Responses (RIRs) in the environment where the recognizer is to be used. In some important applications, measuring a set of RIRs in the target room is either not possible or too expensive. An alternative approach based on Maximum Likelihood (ML) estimation using the reverberant feature sequences of a few calibration utterances with known transcriptions [5] is able to capture the effect of the early reflections [6] in the RIR very well, but overestimates the effect of the late reflections.

Therefore, we propose a combined approach to estimate the RVM based on the recordings of a few calibration utterances in the target environment. In the proposed approach, the RVM is obtained by determining the early, the late and the single-slope decay rates which will be defined in Sec. 3. The ML approach of [5] is used to determine the early decay rate of the RIR in each mel channel. Then, the single-slope decay rate in each mel channel is estimated by a blind method for the determination of reverberation time based on the distribution of signal decay rates [7]. Using the estimates for the early and the single-slope decay rate, the late decay rate in each mel channel is determined by adjustment of the single-slope decay rate. Thus, an RVM capturing both the initial and the late reverberation with relatively high accuracy is obtained.

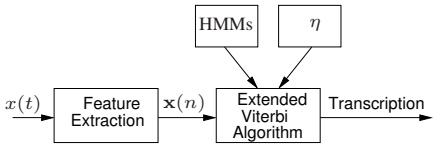

Fig. 1. Structure of a REMOS-based recognizer according to [1].

The paper is structured as follows: The underlying algorithms are concisely reviewed in Sec. 2 to prepare the description of the combined approach in Sec. 3. The performance of the proposed approach is evaluated by connected digit recognition experiments in Sec. 4, and conclusions are drawn in Sec. 5.

2. REVIEW OF UNDERLYING ALGORITHMS

2.1. The REMOS Concept

The REMOS concept [1] uses an acoustic model consisting of a clean-speech HMM network and a statistical RVM η for speech recognition as illustrated in Fig. 1. In the mel-frequency spectral (melspec) domain, the clean-speech HMM output sequence $\mathbf{s}(n)$ and the output sequence $\mathbf{h}(m,n)$ of the reverberation model can be combined by a feature-domain convolution in order to describe the reverberant feature vector sequence $\mathbf{x}(n)$ [1]. In $\mathbf{h}(m,n)$, n is the observation frame index, and m is the reverberation frame index.

The RVM η can be considered as a feature-domain representation of the RIR. As in real-world applications the RIR is usually unknown and time-varying, a fixed feature-domain RIR representation is not sufficient to describe the reverberation. Instead, a statistical RVM η is introduced in [1]. The RVM exhibits a matrix structure where each row corresponds to a certain mel channel and each column to a certain frame as shown in Fig. 2. Each matrix element is modeled by a Gaussian Independent Identically Distributed (IID) random processes. For simplicity, the elements are assumed to be mutually statistically independent [1]. Thus, the RVM is completely described by its mean matrix $m_{\mathbf{H}}$ and its variance matrix $\sigma^2_{\mathbf{H}}$.

For recognition, an extended version of the Viterbi algorithm is employed [1] to find the most likely path through the network of HMMs. The reverberation model η is taken into account by an inner

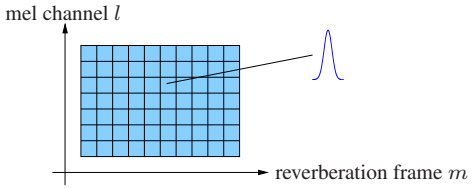

Fig. 2. Reverberation model η for observation frame n.

optimization operation determining the most likely contribution of the current HMM state and the reverberation model to the current reverberant observation vector $\mathbf{x}(n)$. As the inner optimization operation determines the most likely clean-speech feature vector, it is the core of the feature-domain dereverberation algorithm.

2.2. ML Estimation of the RVM

In [5], the feature-domain representation of the RIRs in the target environment is determined by ML estimation based on the reverberant feature sequences of a few calibration utterances with known transcriptions. For each utterance, the reverberant speech signal is segmented into hyper-frames and transformed to the melspec domain. For each hyper-frame, a speech model describing the clean-speech hyper-frame is determined by aligning the HMM sequence of the known transcription to the reverberant utterance. From this clean-speech model, a reverberant speech model is derived by replacing the mean vectors of the clean-speech model with the melspec convolution of the clean means and the unknown melspec RIR representation. Then the likelihood of the reverberant hyper-frame given the reverberant speech feature sequence and the reverberant speech model is maximized with respect to the melspec RIR representation. Thus, a melspec RIR representation is obtained for each hyper-frame. By averaging over the melspec RIR representations of all hyper-frames, the means and the variances of the reverberation model are obtained [5].

2.3. Blind Estimation of Reverberation via Mapping of Statistical Features

In [7], a method for blind estimation of the reverberation time based on the distribution of signal decay rates is presented and its accurate performance for 'diffuse RIRs' is shown. In the context of this paper, a diffuse RIR, or part thereof, is defined where the energy envelope exhibits a single exponential decay. Such an RIR can be described in the Short-Time Fourier Transform (STFT) domain by

$$\ln \tilde{H}(m,k) = \ln P(k) + \lambda_h(k)\, m \quad \text{for} \quad m \geq 0, \tag{1}$$

where $\tilde{H}(m,k)$ is the energy envelope of the RIR at (reverberation) frame m and frequency bin k, $\lambda_h(k)$ is the decay rate, and $P(k)$ is the frequency response of the initial reverberation frame. Based on this model, the decay rate $\lambda_h(k)$ can be estimated by applying a linear fit to the logarithm of the time-frequency energy envelope. In the following discussion, the frequency-dependence is dropped for clarity.

The estimated probability density function (pdf) of the decay rate of a reverberant speech signal becomes increasingly 'skewed' as the decay rate decreases (or equivalently as the reverberation time T_{60} increases) [7]. Thus, the 'skewness' of the estimated pdf can be used to estimate the decay rate of the RIR envelope. As a measure for the 'skewness', the negative-side variance σ^2_{X-} is proposed in [7] because of its superior properties compared to the third central moment. The negative-side variance σ^2_{X-} of a random variable X is

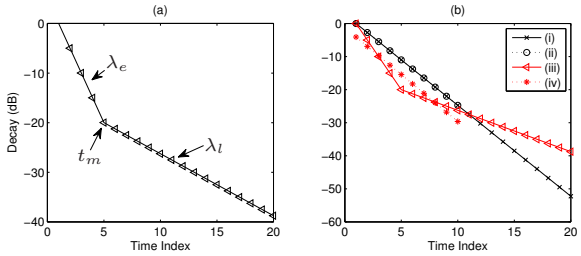

Fig. 3. (a) Two-slope decay model and its parameters. (b) Examples of single-slope estimations on a two-slope decay model.

defined as the variance corresponding to a symmetrical probability density function (pdf) $f^-_X(\lambda)$ with the same negative-side pdf as the original pdf $f_X(\lambda)$ according to

$$f^-_X(\lambda) = \begin{cases} f_X(\lambda) & \text{for } \lambda \leq 0, \\ f_X(-\lambda) & \text{if } \lambda > 0. \end{cases} \tag{2}$$

A second-order function is used in [7] to map the observed σ^2_{X-}, obtained from the reverberant speech decay rate distribution, to the estimated true room decay rate $\hat{\lambda}_h$ as

$$\hat{\lambda}_h = \gamma_2(\sigma^2_{X-})^2 + \gamma_1\sigma^2_{x-} + \gamma_0. \tag{3}$$

The parameters $(\gamma_0, \gamma_1, \gamma_2)$ of the mapping function are obtained in [7] by using Polack's statistical reverberation model [8] and two speech fragments consisting of one male and one female sentence. It should be noted that the parameters depend on the STFT and the decay rate fitting implementations. Since this is a single-slope estimation, it will underestimate the late reverberation for non-diffuse RIRs. In the general non-diffuse case, it is not possible to avoid this underestimation using a single-slope estimation without selective slope fitting such as the method described by Lebart et al. [9].

3. COMBINED APPROACH

In this paper, we present a two-slope RIR model for non-diffuse RIRs. The discrete-time two-slope decay logarithmic envelope $D(t)$ is extended from Pollack's time-domain model [8] to

$$\begin{aligned} D(t) &= \lambda_e t && \text{for} \quad 0 \leq t \leq t_m, \\ D(t) &= (\lambda_e - \lambda_l)t_m + \lambda_l t && \text{for} \quad t > t_m. \end{aligned} \tag{4}$$

where λ_e is the decay of the early reflections arriving before the 'mixing time', t_m, and λ_l is the decay of the late reverberation originating from the diffuse field. Fig. 3(a) illustrates a two-slope decay room model and its parameters.

3.1. Late Decay Adjustment

The curves (i) and (ii) in Fig. 3(b) represent a diffuse room model and its single-slope estimate, respectively. Since the single-slope estimate is able to perfectly capture the envelope of the diffuse RIR, the curves (i) and (ii) are virtually identical. However, for the envelope of a non-diffuse RIR as illustrated in Fig. 3(b) curve (iii), the single-slope estimate of curve (iv) provides only a basic approximation. In the following section, we will derive an approach for the determination of the late decay rate λ_l given estimates for a single-slope decay and an early decay rate λ_e. Single-slope estimation using linear least squares i.e. $v = \alpha u + \beta$, on a two-slope model can be expressed as

$$[\hat{\alpha}\ \hat{\beta}]^T = (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{v}, \tag{5}$$

where $\mathbf{U}$ is the matrix of the independent variable (time index), i.e.

$$\mathbf{U} = \begin{bmatrix} 1 & 2 & 3 & \dots & N \\ 1 & 1 & 1 & \dots & 1 \end{bmatrix}^T, \quad (6)$$

N is the number of observations, and $\mathbf{v}$ is the vector notation for the values of the two-slope RIR model given by (4)

$$\mathbf{v} = [\mathbf{r}\ (\lambda_e - \lambda_l)\mathbf{t}_m + \mathbf{q}]^T. \quad (7)$$

Here, $\mathbf{r} = [0\ \lambda_e\ 2\lambda_e \dots t_m\lambda_e]$, $\mathbf{t}_m = [t_m\ t_m \dots]$ and $\mathbf{q} = [\lambda_l\ 2\lambda_l\ \dots]$. Evaluating the first row of (5), the single-slope estimate $\hat{\alpha}$ can be written as $\hat{\alpha} = \mathbf{a}\mathbf{v}$, where

$$\mathbf{a}_i = \frac{12}{N^3 - N} \cdot \left[i - \frac{N+1}{2} \right], \quad i \in \{1, 2, \dots, N\} \quad (8)$$

is obtained by calculating the first row of the pseudo-inverse in (5). Evaluating the scalar product $\mathbf{a}\mathbf{v}$, the slope $\hat{\alpha}$ can be written as

$$\hat{\alpha} = \gamma\lambda_e g_1(N, t_m) - \gamma\lambda_l g_2(N, t_m), \quad (9)$$

where $\gamma = (N^3 - N)$ and

$$g_1(N, t_m) = -t_m(t_m - 1)(2t_m - 1 - 3N) \quad (10)$$

$$g_2(N, t_m) = -(2t_m - 1 + N)(t - N)(t - 1 - N). \quad (11)$$

The mixing time t_m is assumed to be 50 ms [6] so that the values of g_1, g_2 and γ can be pre-calculated. Solving (9) for λ_l yields the following estimate for the late decay rate

$$\hat{\lambda}_l = \frac{\hat{\lambda}_e g_1(N, \hat{t}_m) - \hat{\alpha}/\gamma}{g_2(N, \hat{t}_m)}. \quad (12)$$

Note that the above derivation can be extended to logarithmic magnitudes in the STFT domain.

3.2. Implementation

We first estimate an STFT-domain representation $H(m, k)$ of the RIR using the single-slope method of [7]. To increase the robustness of the frequency-dependent decay estimates from the single-slope estimation, a rectangular window is used to smooth across the frequency bins on both the calibration and the estimation stage. Transforming $H(m, k)$ to the melspec domain, we obtain the melspec RIR representation $H^*_{\text{mel}}(m, l)$, where l is the mel channel index.

Since the ML approach of [5] captures the early decay very well, $\hat{\lambda}_e$ is obtained by performing a linear fit on the first 50 ms from the ML estimate. Based on the estimates $\hat{\lambda}_e$ and $\hat{\alpha}$, for each mel channel, the late decay adjustment is carried out according to (12). Since both $\hat{\lambda}_e$ and $\hat{\alpha}$ are estimated, and t_m is assumed constant, a particular adjustment may exhibit a significant estimation error. Therefore, each adjusted late decay is smoothed according to

$$\hat{\alpha}'_l = \xi_1\hat{\alpha}_q + (1 - \xi_1)\mathrm{E}[\hat{\alpha}_l]_l, \quad (13)$$

where ξ_1 is the first-stage decay smoothing parameter and $\mathrm{E}[]_l$ denotes the expectation operator across the mel channels. A raw adjusted melspec RIR representation $H^{(1)}_{\text{mel}}(m, l)$ is then generated using $\hat{\lambda}_e$ and $\hat{\alpha}'_l$. Smoothing $H^{(1)}_{\text{mel}}(m, l)$, an improved melspec RIR representation is obtained as

$$H^{(2)}_{\text{mel}}(m, l) = \xi_2 H^{(1)}_{\text{mel}}(m, l) + (1 - \xi_2)\mathrm{E}[H^{(1)}_{\text{mel}}(m, l)]_l, \quad (14)$$

where ξ_2 is the second-stage smoothing parameter. The mean matrix $m_{\mathbf{H}}$ of the REMOS reverberation model is calculated by averaging over the estimates $H^{(2)}_{\text{mel}}(m, l)$ obtained for several utterances. For the estimation of the variance matrix $\sigma^2_{\mathbf{H}}$, a heuristic approach is used. Comparing the mean matrix and the variance matrix of the RVMs according to [1], it is observed that $\sigma^2_{\mathbf{H}}$ is very close to $m^2_{\mathbf{H}}$, where the superscript denotes element-wise squaring. Therefore, the variance matrix is obtained by calculating the element-wise square of the mean matrix $\sigma^2_{\mathbf{H}} = m^2_{\mathbf{H}}$ in the proposed approach.

4. EXPERIMENTS

Experiments with the same connected-digit recognition task as used in [1, 5] are carried out to analyze the performance of the reverberation models determined according to Sec. 3 in connection with the REMOS concept.

4.1. Experimental Setup

In real-world applications, the proposed approach can be used as follows. If the recognizer is to be used in a new room, a few calibration utterances with known transcriptions have to be recorded by the recognizer's distant-talking microphone. Due to the low complexity of the proposed combined approach, the RVM for the corresponding room can then be calculated within a few seconds, and the REMOS-based recognizer is ready for operation. The following experimental setup aims at simulating this real-world scenario as accurately as possible.

The experimental setup is identical to that of [5]. Therefore, only the most important facts are recalled here. Static melspec features with 24 mel channels calculated from speech data sampled at 20 kHz are used. 16-state word-level HMMs with single Gaussian densities serve as clean-speech models. To obtain the reverberant test data (and the reverberant training data for the training of reverberant HMMs used for comparison), the clean-speech TI digits data are convolved with different RIRs measured at different loudspeaker and microphone positions in three rooms with the characteristics given in Table 1. Each test utterance is convolved with an RIR selected randomly from a number of measured RIRs in order to simulate changes of the RIR during recognition.

For the ML estimation of the reverberation models according to [5], 20 calibration utterances from the TI digits training set are convolved with the measured RIRs from the training set and transformed to the feature domain. Each utterance is used as one hyper-frame so that the ML estimate is based on an average over 20 hyper-frames.

To maintain a strict separation of training data (speech and RIRs) from the test data in all experiments, the RIRs from room C are used as calibration set while the tests are performed in room A and B. Comparing the closeness of the melspec RIR representation $H^{(2)}_{\text{mel}}(m, l)$ to the mean matrix of the exact RVM according to [1] for room C, the smoothing parameters ξ_1 and ξ_2 were chosen as 0.5 and 0.5, respectively, for a trade-off between frequency characteristics capture and outlier robustness. The estimates $H^{(2)}_{\text{mel}}(m, l)$ are calculated for 7 calibration utterances so that the mean matrix $m_{\mathbf{H}}$ is obtained by averaging over 7 different estimates $H^{(2)}_{\text{mel}}(m, l)$.

4.2. Experimental Results

Fig. 4 compares the means of the RVMs for room B obtained by (a) measuring RIRs according to [1], (b) ML estimation according to [5], and (c) the proposed combined approach according to Sec. 3.

	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
M	20	50	70

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

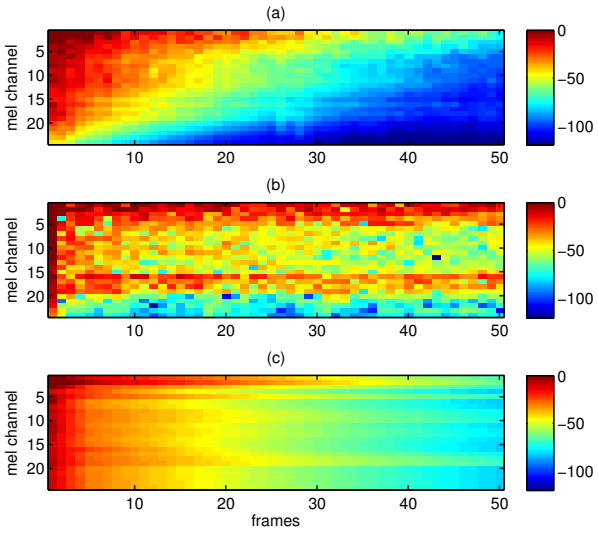

Fig. 4. Mean matrix $m_{\mathbf{H}}$ of the RVM for Room B: (a) exact RVM according to [1], (b) ML-RVM according to [5], and (c) combined approach according to Sec. 3.

While the ML estimate (b) significantly overestimates the late reverberation, this part is captured with relatively high accuracy by the proposed approach (c). Only the strong low-pass characteristic of the room transfer function is not precisely modeled by (c).

Table 2 compares the word accuracies of conventional HMM-based recognizers to that of the REMOS concept using RVMs estimated by different algorithms. The RVMs determined according to Sec. 3 (V) outperform the ML-RMVs (IV) and the HMM-based recognizer trained on matched reverberant data (II) in both rooms since they also capture the effect of late reverberation with relatively high accuracy. In room A, the performance of the RVM (V) even approaches that of the exact RVM (III). The gain of (V) compared to (II) and (IV) is somewhat lower in room B, since it is relatively difficult to capture the strong low-pass characteristic of room B with the proposed combined approach.

5. SUMMARY AND CONCLUSIONS

A combined approach for the estimation of feature-domain reverberation models for robust distant-talking ASR based on the REMOS concept [1] has been proposed in this paper. Using only a few calibration utterances recorded in the target environment, the proposed approach determines the means and variances of a matrix-valued IID Gaussian random process. The initial frequency response and the early decay is determined by ML estimation according to [5]. Blind estimates of the reverberation time according to [7] are used to determine single-slope decay estimates. These single-slope estimates include the early decay and the late decay. Using the estimates for the early and the single-slope decay rate, the late decay rate in each mel channel is determined by adjustment of the single-slope decay rate. Thus, an RVM capturing both the initial and the late reverberation with high accuracy is obtained. Since the parameters of the RVM are estimated without the need for close-talking recordings or RIR measurements, the effort for training is reduced compared to the estimation method proposed in [1], and it is greatly reduced compared to the training of HMMs on artificially reverberated data. Simulation results of a connected digit recognition task confirm that using the reverberation models obtained by the proposed combined approach in the REMOS concept significantly outperforms the reverberation models based on ML estimation [5] and also conventional HMM-based recognizers trained on matched reverberant data.

	clean data	Room A	Room B
(I) conventional HMMs, clean training	82.0	51.5	13.4
(II) conventional HMMs, reverberant training	-	66.8	54.6
(III) REMOS, exact RVM according to [1]	-	77.6	71.6
(IV) REMOS, ML-RVM according to [5]	-	63.0	57.3
(V) REMOS, RVM according to Sec. 3	-	74.5	60.4

Table 2. Word accuracies for the conventional HMM-based recognizer trained on clean (I) and reverberant speech (II) and for the REMOS concept [1] using exact RVMs (III), ML RVMs according to [5](IV), and RVMs according to Sec. 3 (V).

6. 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. INTERSPEECH*, pp. 769–772, 2006.

[2] 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.

[3] P. A. Naylor and N. D. Gaubitch, "Speech dereverberation," *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, September 2005.

[4] H.-G. Hirsch and H. Finster, "A new approach for the adaptation of HMMs to reverberation and background noise," *Speech Communication*, vol. 50, no. 3, pp. 244–263, March 2008.

[5] A. Sehr, Y. Zheng, E. Nöth, and W. Kellermann, "Maximum likelihood estimation of a reverberation model for robust distant-talking speech recognition," *Proc. European Signal Processing Conference (EUSIPCO)*, pp. 1299–1303, 2007.

[6] H. Kuttruff, *Room Acoustics*, 4th ed. Taylor & Francis, Oct. 2000.

[7] J. Y. C. Wen, E. A. P. Habets, and P. A. Naylor, "Blind estimation of reverberation time based on the distribution of signal decay rates," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 329–332, March 2008.

[8] J. D. Polack, "La transmission de l'énergie sonore dans les salles," Ph.D. dissertation, Université du Maine, Le Mans, 1988.

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