BLIND ESTIMATION OF A FEATURE-DOMAIN REVERBERATION MODEL IN NON-DIFFUSE ENVIRONMENTS WITH VARIANCE ADJUSTMENT

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

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

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

ABSTRACT

Blind estimation of a two-slope feature-domain reverberation model is proposed. The reverberation model is suitable for robust distant-talking automatic speech recognition approaches which use a convolution in the feature domain to characterize the reverberant feature vector sequence, e.g. [1, 2, 3]. Since the model describes the reverberation by a matrix-valued IID Gaussian random process, its statistical properties are completely captured by its mean and variance matrices. The suggested solution for the estimation of the model includes two novel features based on the study of simulated rooms: 1) a solution for blindly determining a two-slope decay model from a single-slope estimate; 2) a variance mask to improve the estimation of the variance matrix. Using the proposed solution, the reverberation model can be estimated during recognition without the need of pre-training or using calibration utterances with known transcription. Connected digit recognition experiments using [3] show that the reverberation models estimated by the proposed approach significantly outperform HMM-based recognizers trained on reverberant data in most environments.

1. INTRODUCTION

Distant-talking speech capture can increase the comfort and the acceptance of many Automatic Speech Recognition (ASR) applications, like e.g. automatic meeting transcription, voice control of consumer electronics, and dictation systems. However, the reverberation caused by multi-path propagation of sound waves from the source to the distant-talking microphone leads to a mismatch between the input utterances and the acoustic model of the recognizer, usually trained on close-talking speech. Therefore, the performance of ASR systems is significantly reduced by reverberation [4, 5] if no countermeasures are taken.

In the time domain, reverberant speech can be described by a convolution of clean speech with the Room Impulse Response (RIR) characterizing the acoustic path from the speaker to the microphone. The length of the RIR, typically ranging from 200 ms to 1000 ms, significantly exceeds the length of the analysis window used for feature extraction in ASR systems, typically ranging from 10 ms to 40 ms. Therefore, the time-domain convolution is not transformed into a simple multiplication in the short-time frequency transform (STFT) domain. Instead, reverberation still has a dispersive effect in the STFT domain and also in STFT-based feature domains. To capture this dispersive effect, a convolution of the clean-speech feature vectors with a feature-domain reverberation representation in the mel-spectral (melspec) domain has been proposed in several recent publications, e.g. [1, 2, 3].

Blind estimation of a statistical feature-domain ReVerberation Model (RVM) [3] which can be employed in virtually all robust distant-talking ASR concepts based on the melspec convolution described above, is proposed. Since the RVM describes the reverberation by a matrix-valued IID Gaussian random process, its statistical properties are completely captured by its mean and variance matrices. While a set of known RIRs in [3], simultaneous recordings of close-talking and distant-talking microphones in [6], and calibration utterances with known transcriptions in [1, 7, 8] are required for estimating the reverberation representation, the proposed approach can estimate the RVM blindly during recognition. Thus, the flexibility of the robust distant talking ASR approaches according to [1, 2, 3] can be significantly improved.

The suggested solution includes two new features based on the study of simulated rooms: firstly, a blind solution for determining a two-slope decay model from a single-slope estimate; secondly, a variance mask to improve the estimation of the RVM's variance matrix. Using the proposed solution, the reverberation model can be estimated during recognition without the need of pre-training or using transcribed calibration utterances. The paper is structured as follows: The underlying algorithms are concisely reviewed in Sec. 2 followed by the description of the blind approach and the variance mask in Sec. 3 and Sec. 4, respectively. The performance of the proposed approach is evaluated by connected digit recognition experiments based on the concept of [3] in Sec. 5, and conclusions are drawn in Sec. 6.

2. REVIEW OF UNDERLYING ALGORITHMS

2.1 Statistical RVM

The statistical RVM η used in this contribution has been introduced in [3]. It can be considered as a feature-domain representation of all possible RIRs for arbitrary speaker and microphone positions in a certain room. 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. 1(a). Each matrix element is modeled by a Gaussian Independent Identically Distributed (IID) random process. For simplicity, the elements are assumed to be mutually statistically independent [3]. Thus, the RVM is completely described by its mean matrix $m_{\mathbf{H}}$ and its variance matrix $\sigma^2_{\mathbf{H}}$.

2.2 Blind Estimation of Reverberation

In [9], a method for blind estimation of reverberation time based on the distribution of signal decay rates is presented and its accurate performance for 'diffuse RIRs', that is, for RIRs exhibiting a single exponential decay, is shown. The

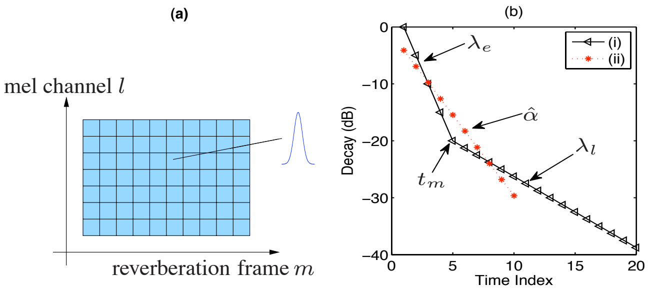

Figure 1: (a) Reverberation model η for observation frame n. (b) Two-slope decay model (i) and its single-slope estimate (ii).

decay rate is defined as the gradient of the first order linear least squares fit.

The estimated probability density function (pdf) of the decay rate of a reverberant speech signal in the STFT log-magnitude domain, λ_x, becomes increasingly 'skewed' as the decay rate decreases or equivalently as the reverberation time T_{60} increases [9]. 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' of the random variable λ_x, the negative-side variance σ^2_{X-} is proposed in [9] because of its superior properties compared to the third-order central moment.

A second-order function is used in [9] to map the observed σ^2_{X-}, obtained from the reverberant speech decay rate distribution, to the estimated single-slope room decay rate $\hat{\alpha}$. In this contribution, a more generic mapping function for the estimation of the room decay $\hat{\alpha}$ according to

$$\hat{\alpha} = \sum_{r=0}^{4} \gamma_r (\sigma^2_{X-})^r \qquad (1)$$

is used. The parameters (γ_r) of the mapping function are obtained in [9] by using Polack's statistical reverberation model [10] and two speech fragments consisting of one male and one female sentence.

2.3 Late Decay Adjustment

RIRs obtained in real-world rooms are not 'diffuse' since 'diffuse' RIRs require an infinite source-microphone distance. Non-perfectly diffuse sound fields exhibit a faster decay for the early segment corresponding to the direct sound and early reflections, and a slower decay for the late reverberation [11, 12]. Therefore, a two-slope RVM extended from Pollack's time-domain model [10] is used in [8] to capture the non-diffuse RIRs as depicted in Fig. 1(b). In the early segment of the two-slope model, extending from time index one to the mixing time t_m, the envelope decreases with the early decay rate λ_e. In the late segment, starting at t_m, the envelope decreases with the late decay rate λ_l. Given an estimate for the single-slope decay, $\hat{\alpha}$, determined according to [9] and an estimate for the early decay $\hat{\lambda}_e$ determined from the ML estimate according to [7], an estimate of the late decay $\hat{\lambda}_l$ is obtained by linear least squares adjustment in [8].

3. BLIND DECAY ADJUSTMENT

Since the late decay adjustment according to [8] described above uses the ML approach of [7] for estimating the early decay rate, calibration utterances with known transcription are necessary for the determination of the RVM. In this section, a blind approach for the adjustment of the late decay is introduced which does not require pre-transcribed calibration utterances.

Let $\hat{\lambda}_e^*$ be the early decay rate when the source-mic distance is greater than the critical distance, where the superscript * indicates that the microphone is located in the diffuse sound field. In this case, a fixed ratio between $\hat{\lambda}_e^*$ and $\hat{\lambda}_l$ is assumed for simplicity. Thus, the single-slope decay can be estimated by linear least squares optimization from N observations as (see [8] for a derivation)

$$\hat{\alpha}_{\max} \approx \gamma \hat{\lambda}_e^* g_1(N,t_m) - \gamma \hat{\lambda}_l g_2(N,t_m), \qquad (2)$$

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

$$\begin{aligned} g_1(N,t_m) &= -t_m(t_m-1)(2t_m-1-3N) \qquad (3)\\ g_2(N,t_m) &= -(2t_m-1+N)(t_m-N)(t_m-1-N). \qquad (4) \end{aligned}$$

The mixing time t_m is assumed to be 50 ms [13] so that the values of g_1, g_2 and γ can be pre-calculated. If we assume a fixed ratio such that $\hat{\lambda}_e^* = \varpi \hat{\lambda}_l$, where ϖ is determined in Section 4, we can estimate $\hat{\lambda}_l$ as

$$\hat{\lambda}_l \approx \frac{\hat{\alpha}_{\max}}{\varpi \gamma g_1(N,t_m) - \gamma g_2(N,t_m)}. \qquad (5)$$

We first estimate an STFT-domain representation $H^{(1)}(m,k)$ of the RIR (m being the reverberation frame index and k indexing the frequency bins) using the single-slope method of [9]. To increase the robustness of the frequency-dependent decay estimates, a rectangular window is used to smooth across the frequency bins k of $H^{(1)}(m,k)$. Then the early and late decays are estimated using (5) to obtain a two-slope adjustment $H^{(2)}(m,k)$ in the STFT-domain. Since $\hat{\alpha}$ is an estimate, and both t_m and ϖ are assumed constant, a particular adjustment may exhibit a significant estimation error. Therefore, smoothing across k is performed according to

$$H^{(3)}(m,k) = \xi H(^{(2)}m,k) + (1-\xi)\mathrm{E}[H^{(2)}(m,k)]_k \qquad (6)$$

to get a smoothed STFT-domain RIR representation, where ξ is the smoothing parameter and $\mathrm{E}[]_k$ denotes the expectation across frequency bins k. Transforming $H^{(3)}(m,k)$ to the melspec domain, we obtain the melspec RIR representation $H_{\mathrm{mel}}(m,l)$, where l is the mel channel index. The mean matrix $m_{\mathbf{H}}$ of the RVM is calculated by averaging over the estimates $H_{\mathrm{mel}}(m,l)$ obtained for several utterances.

For the estimation of the variance matrix $\sigma^2_{\mathbf{H}}$, a heuristic approach is used in [8]. Based on a comparison of the mean matrix and the variance matrix of the RVMs according to [3], an estimate $\tilde{\sigma}^2_{\mathbf{H}} = m^2_{\mathbf{H}}$ of the variance matrix is obtained by calculating the element-wise square of the mean matrix.

4. VARIANCE MASK

An in-depth investigation of the relationship between the heuristic variance estimate $\tilde{\sigma}^2_{\mathbf{H}} = m^2_{\mathbf{H}}$ according to [8] and the reference variance estimate $\sigma^2_{\mathbf{H}}$ according to [3] shows some systematic dependencies which can be used to improve the estimate $\tilde{\sigma}^2_{\mathbf{H}}$. Therefore, we propose a variance mask, $\varsigma^2_{\mathbf{H}}$, to map $\tilde{\sigma}^2_{\mathbf{H}}$ as close as possible to $\sigma^2_{\mathbf{H}}$ in the form of

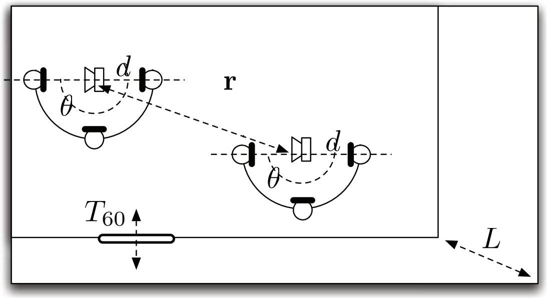

Figure 2: Simulation setup of different room acoustics parameters using the image method of [14].

$$\hat{\sigma}^2_{\mathbf{H}} \doteq \varsigma^2_{\mathbf{H}} \otimes \tilde{\sigma}^2_{\mathbf{H}} = \varsigma^2_{\mathbf{H}} \otimes m^2_{\mathbf{H}} \overset{!}{\approx} \sigma^2_{\mathbf{H}} \tag{7}$$

where $\otimes$ denotes the Hadamard product. Taking the natural logarithm of each matrix element, we get

$$\ln \hat{\sigma}^2_{\mathbf{H}} = \ln \varsigma^2_{\mathbf{H}} + \ln m^2_{\mathbf{H}}. \tag{8}$$

To determine the parameters of the variance mask, the relations between the characteristics of the variance mask and room acoustic properties are studied. The influence of the following acoustic parameters is investigated: i) source-mic distance $d = 1 \ldots 4$ m, ii) reverberation time $T_{60} = 0.2 \ldots 1.0$ s, iii) room size $L = 120 \ldots 320$ m^3 and iv) source-microphone position $\mathbf{r}$ relative to the room. For each of the parameters ii), iii), and iv), a set of RIRs with fixed parameter i) on a semi-circle as illustrated in Fig. 2 is generated using the image method [14]. Each set consists of 20 RIRs determined for different microphone positions (θ) on a semi-circle. Based on the sets of RIRs, the reference values of $\sigma^2_{\mathbf{H}}$ and $m^2_{\mathbf{H}}$ are calculated according to [3].

Two observations are made regarding the characteristics of the variance mask for the individual matrix elements (m,l): Firstly, $\tilde{\sigma}^2_H(m,l)$ overestimates the reference variance $\sigma^2_H(m,l)$ in the first frame $m = 1$, i.e., the variance mask should be negative for $m = 1$. The overestimation is mainly due to the nearly constant direct component dominating the first frame. Secondly, there is an increasing overestimation of the reference variance $\sigma^2_H(m,l)$ by $\tilde{\sigma}^2_H(m,l)$ for increasing mel-channel index, i.e., the variance mask should be decreasing with increasing frequency. A possible explanation for this observation could be that for increasing frequency, the density of the normal modes increases according to statistical room acoustics [13]. Since this means averaging over a higher number of modes for high frequencies, the variance of the feature domain RIR representation due to position changes decreases with frequency. Therefore, we propose the following variance mask with ρ and φ as parameters:

$$\ln \varsigma^2_H(m,l) = \begin{cases} \rho & \text{for } m = 1\ ,\ l = 1,\ldots,24, \\ 0 & \text{for } m > 1, l = 1 \ldots 4, \\ \varphi(m-4) & \text{for } m > 1, l = 5 \ldots 24. \end{cases} \tag{9}$$

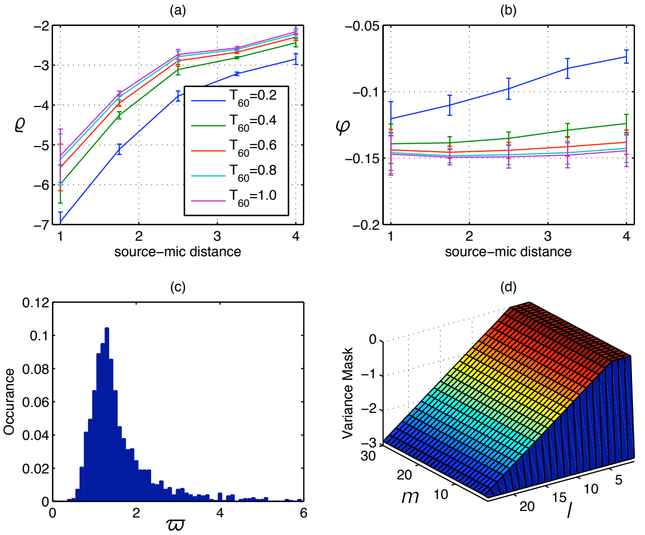

Figure 3: Mean optimised parameters (a) ρ and (b) φ for different source-mic distances and different reverberation times for the two-parameter variance mask model. (c) Histogram of ϖ averaged over all frequencies. (d) Variance mask $\ln \varsigma^2_{\mathbf{H}}$ for optimal ρ and φ.

The proposed variance mask has the shape characteristics shown in Fig. 3(d). Since the image method is known to be relatively inaccurate for lower frequencies [14], we set the first four mel channels of the variance mask $\ln \varsigma^2_{\mathbf{H}}$ to zero. Minimising the error between the reference $\sigma^2_{\mathbf{H}}$ and the estimate $\hat{\sigma}^2_{\mathbf{H}}$ based on the two-parameter variance mask with respect to ρ and φ according to

$$\min_{\rho,\varphi} \{ \| \sigma^2_{\mathbf{H}} - \hat{\sigma}^2_{\mathbf{H}} \|_2 \}, \tag{10}$$

where $\| \cdot \|_2$ denotes the spectral norm, we obtain the parameters ρ and φ. In Fig. 3 (a) and (b), optimised parameters ρ and φ for different source-microphone distances and reverberation times are shown. For each source-microphone distance and reverberation time, the mean and the variance across five different room sizes and five relative positions are determined. We see that the source-mic distance dominates the term ρ while the reverberation time only slightly affects it. The variation of the parameter φ with respect to the source-mic distance increases with increasing reverberation time.

For the intended applications, like meeting transcription or voice control of consumer electronics, we assume an average reverberation time of $T_{60} = 0.6$ s and an average source-mic distance of $d = 2.5$ m to select the optimum values of the parameters ρ and φ. Connected digit recognition tests in [15] indicate that overestimation of the reverberation by the RVM is less detrimental than underestimation. Since larger values of ϖ cause the adjusted late decay to be slower, we select $\varpi = 3$ corresponding to the 0.9-percentile of the ϖ values found in the image method rooms shown in Fig. 3(c). A hard decision at the 0.9-percentile ensures most of the decays are adjusted with slight overestimation of the late decay, and only a few decays with underestimation.

5. EXPERIMENTS

Experiments with the same connected-digit recognition task as used in [3, 7, 8] are carried out to analyze the performance of the reverberation models determined according to Sec. 3 and 4. For recognition, the approach of [3] is used.

5.1 Experimental Setup

In real-world applications, the proposed approach can be used as follows. If the recognizer is to be used in a new

(a)	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 m
SRR	4.0 dB	-4.0 dB	-4.0 dB
M	20	50	70

▨ - Variance Mask

▨ - Blind Approach

(b)	Clean Data	Room								
		A			B			C		
Variance Model	N/A	$\sigma^2_{\mathbf{H}}$	$\tilde{\sigma}^2_{\mathbf{H}}$	$\hat{\sigma}^2_{\mathbf{H}}$	$\sigma^2_{\mathbf{H}}$	$\tilde{\sigma}^2_{\mathbf{H}}$	$\hat{\sigma}^2_{\mathbf{H}}$	$\sigma^2_{\mathbf{H}}$	$\tilde{\sigma}^2_{\mathbf{H}}$	$\hat{\sigma}^2_{\mathbf{H}}$
I	82.0	51.5	-	-	13.4	-	-	25.9	-	-
II	-	66.8	-	-	54.6	-	-	46.0	-	-
III	-	77.6	66.3	***77.5***	71.6	71.6	***75.0***	67.6	63.4	***69.0***
IV	-	76.4	62.9	***67.2***	54.9	31.4	***32.0***	60.6	39.4	***45.3***
V	-	75.9	63.0	***75.8***	62.2	57.3	***71.7***	35.1	58.5	***61.7***
VI	-	78.1	74.5	***78.3***	67.7	60.4	***68.1***	67.7	56.7	***61.0***
VII	-	76.1	67.4	***73.5***	56.5	44.8	***49.3***	62.7	53.9	***59.8***

Table 1: (a) Room Characteristics. (b) Word accuracies in % for the conventional HMM-based recognizer trained on (I) clean and (II) reverberant speech. The ASR concept of [3] with the RVMs estimated according to (III) [3],(IV) single-slope approach, (V) [7], (VI) [8] and (VII) Sec. 3. *Three* variance models are used in connection with the *five* different $m_{\mathbf{H}}$ estimation methods (III)-(VII): $\sigma^2_{\mathbf{H}}$-variance of the RVM is estimated according to [3]; $\tilde{\sigma}^2_{\mathbf{H}}$-using $m^2_{\mathbf{H}}$; $\hat{\sigma}^2_{\mathbf{H}}$- according to Sec. 4 (7).

room, the first utterance to be recognized is used to determine a melspec RIR representation according to Sec. 3. An initial estimate of the RVM is obtained from this single RIR representation, and the recognition is started. As soon as the next utterances are available, they are used to estimate melspec RIR representations, which are used to update the initial RVM. For the following tests, the first seven utterances are used for averaging the RVM, then it is fixed for the following 505 test utterances. Alternatively, a recursive smoothing over melspec RIR representations could be used. Thus the algorithm could even adjust the RVM when the user moves to another room.

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, the clean-speech TI digits data are convolved with different RIRs measured at different LS-mic positions in three rooms with the characteristics given in Table 1(a). 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. Before convolution, the RIRs are normalized to have unit energy in the melspec domain. This normalization corresponds to using automatic gain control as preprocessing in the ASR system.

To maintain a strict separation of the training data from the test data in all experiments, RIRs generated with the image method are used for determining the fixed parameters while the tests are performed in room A, B and C (see Table 1(a)). Comparing the closeness of the melspec RIR representation $H_{\text{mel}}(m,l)$ to the mean matrix of the RVM obtained by averaging over the image method RIRs (see Sec. 4), the smoothing parameters ξ were chosen as 0.1 for a trade-off between frequency characteristics capture and outlier robustness.

5.2 Variance Adjustment

Table 1(b) shows the results of the experiments. The results obtained with the variance estimate $\hat{\sigma}^2_{\mathbf{H}}$ based on the mask according to Sec. 4 are highlighted in the table with cyan background colour. Regardless of the estimation method for determining the mean matrix $m_{\mathbf{H}}$, using the variance estimate $\hat{\sigma}^2_{\mathbf{H}}$ significantly improves the word accuracy over the heuristic estimate $\tilde{\sigma}^2_{\mathbf{H}}$. Using $\hat{\sigma}^2_{\mathbf{H}}$ in connection with the means estimated according to [3] (III) even outperforms the variance estimate $\sigma^2_{\mathbf{H}}$ based on measured RIRs in rooms B and C. In general, the gain of applying the variance mask increases with the accuracy of the $m_{\mathbf{H}}$ estimate. For example, comparing the estimation methods (III) and (VI) in room C, there is a significant difference in the word accuracy for the variance estimate $\hat{\sigma}^2_{\mathbf{H}}$, while the variance estimate $\sigma^2_{\mathbf{H}}$ achieves similar word accuracies for the estimation methods (III) and (VI). The difference in word accuracy for $\hat{\sigma}^2_{\mathbf{H}}$ can be attributed to the slightly more inaccurate mean estimate of (VI). The results for the variance estimate $\sigma^2_{\mathbf{H}}$ based on measured RIRs in connection with the RVM estimation approaches (IV) - (VII) are only given for comparison. In real-world applications, they are only available if the RVM estimation method (III) is used. The relatively low accuracy of 35.1% for (V) in connection with $\sigma^2_{\mathbf{H}}$ is due to the mismatch between the variance matrix $\sigma^2_{\mathbf{H}}$ and the mean matrix $m_{\mathbf{H}}$ for this case.

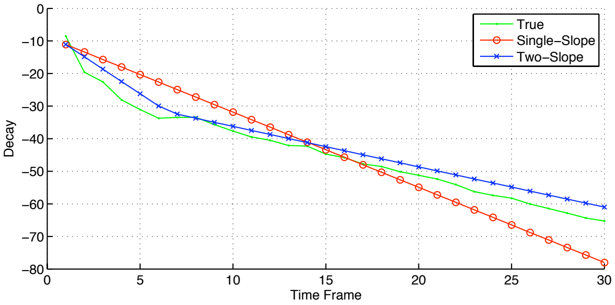

Figure 4: Comparison of the true decay for room C, averaged across all mel channels, with the decay of the corresponding single-slope and two-slope models.

5.3 Blind Estimation

Fig. 4 compares the true decay for room C averaged across all mel channels with the decay of the corresponding single-slope and two-slope models. It is clearly obvious that the single-slope model overestimates the early reverberation and underestimates the late reverberation. The proposed two-slope model is able to capture both early and late parts more accurately. Furthermore, the relatively low recognition rates of initial connected digit recognition tests using the single-slope RVMs in the ASR concept of [3] (see Table1(IV)) confirm that the effect of reverberation cannot be captured with sufficient accuracy by a single-slope model.

The results obtained with the blind adjustment approach (VII) according to Sec. 3 are highlighted by magenta background colour in Table 1(b). For all rooms, the word accuracies obtained by (VII) are significantly higher than that of the conventional HMM-based recognizer trained on clean data (I). In rooms A and C, (VII) also outperforms the conventional HMM-based recognizer trained on matched reverberant data (II). The reason why the performance of (VII) is slightly lower than that of (II) in room B is the strong low-pass characteristic of room B which cannot be perfectly captured by the blind estimation approach (VII). Additional tests with an adjusted RVM where the reference frequency response obtained from measured RIRs is multiplied to the mean matrix $m_{\mathbf{H}}$ show an increase of the word accuracy to 73.8 %. This result indicates that the imperfect capture of the frequency response of room B by (VII) is the main reason for the relatively low recognition rate.

6. SUMMARY AND CONCLUSIONS

Blind estimation of a feature-domain reverberation model for robust distant-talking ASR concepts based on a convolution in the melspec domain, e.g. [1, 2, 3], has been proposed in this paper. The proposed approach determines the mean and the variance matrices of a matrix-valued IID Gaussian random process. Blind estimates of the reverberation time according to [9] are used to determine single-slope decay estimates. Using the proposed adjustment method, the single-slope estimates are transformed to an early and a late decay to produce the mean matrix of a two-slope RVM. The variance matrix of the RVM is determined by the element-wise square of the mean matrix and the proposed variance mask. Thus, an RVM capturing both the initial and the late reverberation as well as the variance with high accuracy is obtained. Since the parameters of the RVM are estimated blindly without the need for close-talking recordings, RIR measurements or transcribed calibration utterances, the RVM can be estimated during recognition so that the recognizer can be used right away. Simulation results of a connected digit recognition task confirm that using the reverberation models obtained by the proposed blind approach in connection with the recognizer concept of [3] achieves similar results as the reverberation models based on non-blind methods [7, 8] and outperforms conventional HMM-based recognizers trained on matched reverberant data in most environments.

REFERENCES

[1] C. K. Raut, T. Nishimoto, and S. Sagayama, “Model adaptation for long convolutional distortion by maximum likelihood based state filtering approach,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Tolouse, France, 2006.

[2] H. G. Hirsch and H. Finster, “A new HMM adaptation approach for the case of a hands-free speech input in reverberant rooms,” in *Proc. Interspeech Conf.*, 2006, pp. 781–783.

[3] A. Sehr, M. Zeller, and W. Kellermann, “Distant-talking continuous speech recognition based on a novel reverberation model in the feature domain,” in *Proc. Interspeech Conf.*, 2006, pp. 769–772.

[4] B. E. D. Kingsbury and N. Morgan, “Recognizing reverberant speech with RASTA-PLP,” *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 1259–1262, 1997.

[5] R. Petrick, K. Lohde, M. Wolff, and R. Hoffmann, “The harming part of room acoustics in automatic speech recognition,” in *Proc. Interspeech Conf.*, 2007, pp. 1094–1097.

[6] C. K. Raut, T. Nishimoto, and S. Sagayama, “Model adaptation by state splitting of HMM for long reverberation,” in *Proc. Interspeech Conf.*, Lisboa, Portugal, September 2005, pp. 277–280.

[7] A. Sehr, Y. Zheng, E. Nöth, and W. Kellermann, “Maximum likelihood estimation of a reverberation model for robust distant-talking speech recognition,” in *Proc. European Signal Processing Conf. (EUSIPCO)*, Poznan, Poland, 2007.

[8] A. Sehr, J. Y. C. Wen, W. Kellermann, and P. A. Naylor, “A combined approach for estimating a feature-domain reverberation model in non-diffuse environments,” in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, 2008.

[9] 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,” in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2008.

[10] J. D. Polack, “La transmission de lénergie sonore dans les salles,” Ph.D. dissertation, Université du Maine, Le Mans, 1988.

[11] M. R. Schroeder, “A new method of measuring reverberation time,” *Journal Acoust. Soc. of America*, vol. 37, no. 3, pp. 409–412, 1965.

[12] P. Kendrick, T. Cox, F. Li, Y. Zhang, and J. Chambers, “Blind estimation of reverberation parameters for non-diffuse rooms,” *Acta Acustica united with Acustica*, vol. 93, no. 5, pp. 760–770, 2007.

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

[14] J. Allen and D. Berkley, “Image method for efficiently simulating small room acoustics,” *Journal Acoust. Soc. of America*, vol. 65, no. 4, pp. 943–950, 1979.

[15] A. Sehr and W. Kellermann, *Speech and Audio Processing in Adverse Environments*. Springer, Berlin, 2008, ch. Towards Robust Distant-Talking Automatic Speech Recognition in Reverberant Environments, pp. 679–728.