A SIMPLIFIED DECODING METHOD FOR A ROBUST DISTANT-TALKING ASR CONCEPT BASED ON FEATURE-DOMAIN DEREVERBERATION

Armin Sehr and *Walter Kellermann*

Multimedia Communications and Signal Processing,
University of Erlangen-Nuremberg
Cauerstr. 7, 91058 Erlangen, Germany
Email: {sehr,wk}@lnt.de

ABSTRACT

A simplified decoding method for the concept of REverberation MOdeling for Speech recognition (REMOS) [1] is proposed. In order to achieve robust distant-talking Automatic Speech Recognition (ASR), the REMOS concept uses a combination of clean-speech HMMs and a reverberation model to perform feature-domain dereverberation during decoding. The simplified decoding/dereverberation method proposed in this contribution significantly reduces the computational complexity of the concept without a major performance reduction.

Index Terms— Dereverberation, robust ASR, reverberation model, feature-domain processing.

1. INTRODUCTION

Distant-talking microphone systems enable human/human and human/machine-interaction without tethering the user to a close-talking microphone. The large distance between speaker and microphone in distant-talking scenarios implies that the microphone does not only pick up the desired signal but also background noise, interfering speakers, and the reverberation of the desired signal caused by multiple reflections of the sound waves at the boundaries of the recording room. The reverberation does not only reduce the perceived sound quality but also decreases the performance of ASR significantly.

The reduced ASR performance is mainly caused by the dispersive effect of reverberation on the feature representation of speech used for ASR. While the feature calculation is based on a short-time spectrum analysis with a typical frame length of 10 to 40 ms, the length of the Room Impulse Response (RIR) describing the acoustic path between speaker and microphone ranges from 200 to 800 ms in typical office or home environments. Therefore, the RIR extends over several frames and the reverberation causes the speech features to be smeared along the time axis. Thus, an effect similar to intersymbol interference known from radio communications is observed.

The dispersion across frames causes the current observed feature vector to depend strongly on the previous feature vectors. Therefore, reverberant feature vector sequences cannot be modeled very well by Hidden Markov Models (HMMs) since one of the basic assumptions underlying HMMs, namely that the current frame depends only on the current state, is heavily violated. Furthermore, the potential gain of model adaptation and compensation techniques working only within one frame, like cepstral mean subtraction [2] or maximum likelihood linear regression [3], is limited.

A promising way to tackle the dispersion problem is to dereverberate the speech signal before feature extraction. Numerous approaches based on linear prediction (e. g. [4, 5]), multi-channel deconvolution (e. g. [6, 7, 8]), and spectral subtraction (e. g. [9, 10]) have been proposed. A recent suggestion to use all available prior knowledge about the source signal in a probabilistic framework [11] appears to be particularly interesting for dereverberation used as preprocessing for ASR. Here, it is possible to employ the extremely powerful speech models of the recognizer to describe the source signal.

The REMOS concept [1], already proposed in 2006, is heading in a similar direction as [11] by combining the clean-speech model represented by the recognizer's HMMs and a statistical reverberation model to perform dereverberation directly in the feature domain. A simplified decoding method for the REMOS concept is proposed in this paper to reduce the computational complexity of the approach.

The paper is structured as follows: Elements of the REMOS concept necessary for the description of the new decoding method in Section 3 are reviewed in Section 2. Experimental results are discussed in Section 4, and conclusions are drawn in Section 5.

2. REVIEW OF THE REMOS CONCEPT

The acoustic model used in the REMOS concept is a combination of a clean-speech HMM network $\mathcal{N}_\lambda$ and a statistical reverberation model η as illustrated in Figure 1. If melfrequency spectral (melspec) coefficients (see Figure 2) are

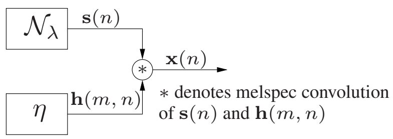

Fig. 1. Acoustic model of the REMOS concept.

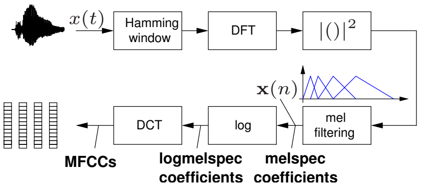

Fig. 2. Calculation of melspec features.

used as features, as assumed throughout the paper, the reverberant feature vector sequence $\mathbf{x}(n)$ can be approximated by a convolution of the clean-speech HMM output sequence $\mathbf{s}(n)$ and the output sequence $\mathbf{h}(m,n)$ of the reverberation model (see [12])

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

Here, $\odot$ denotes element-wise multiplication, M is the length of the reverberation model, and m, n are frame indices. Note that the vector $\mathbf{x}(n) = [x_0(n), \ldots, x_{L-1}(n)]^T$ with T denoting transpose consists of L features $x_l(n)$. The vectors $\mathbf{s}(n)$ and $\mathbf{h}(m,n)$ are defined accordingly.

The reverberation model η represents the RIR in the feature domain. As in real-world applications the RIR is usually unknown and time-varying and the approximation errors of the melspec convolution (1) lead to further variability, a fixed feature-domain RIR representation is not sufficient to describe the reverberation. Instead, a statistical reverberation model η is suggested in [1].

The reverberation model η exhibits a matrix structure where each row corresponds to a certain mel channel and each column to a certain frame as shown in Figure 3. The matrix elements are modeled by Independent Identically Distributed (IID) random processes. For simplicity, the random processes of the different matrix elements are assumed to be statistically independent and normally distributed. Thus, η can be considered as a matrix-valued IID Gaussian random process.

For recognition, an extended version of the Viterbi algorithm is employed [1]. Its recursion equation is given by

$$\gamma_j(n) = \max_i\{\gamma_i(n-1) \cdot a_{ij} \cdot O_{ij}(n)\}, \tag{2}$$

$$O_{ij}(n) = \max_{\mathbf{s}(n),\mathbf{h}(m,n)} \{\, f_\lambda(j,\mathbf{s}(n)) \cdot f_\eta(\mathbf{h}(m,n))\} \quad \text{s.t.} \tag{3}$$

$$\mathbf{x}(n) = \mathbf{h}(0,n) \odot \mathbf{s}(n) + \sum_{m=1}^{M-1} \mathbf{h}(m,n) \odot \check{\mathbf{s}}_{ij}(n-m). \tag{4}$$

Here, $\gamma_j(n)$ is the Viterbi metric for state j at frame n, a_{ij} is the transition probability from state i to state j, $f_\lambda(j,\mathbf{s}(n))$ and $f_\eta(\mathbf{h}(m,n))$ are the output densities of the HMM λ and the reverberation model η, respectively. The term $O_{ij}(n)$ can be considered as the output density of the combined acoustic model. It is calculated by maximizing the joint density of the HMM and the reverberation model subject to (s. t.) the constraint that the combination of the model outputs is equal to the current reverberant observation. This maximization is called inner optimization in the following. The estimated clean-speech vectors $\check{\mathbf{s}}_{ij}(n-m)$ are known from previous iterations of the algorithm. They are retrieved from a matrix of clean speech vectors (3D tensor) which is set up during decoding. The correct clean speech estimates are selected by tracing back the most likely partial path with previous state i and current state j (see [12] for details).

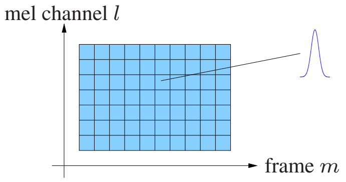

Fig. 3. Reverberation model η for observation frame n.

The inner optimization of Equation (3) s. t. (4) finds the optimum contributions of the HMMs and the reverberation model to the current observation. Since it determines an estimate $\hat{\mathbf{s}}_{ij}(n)$ of the current clean-speech vector, it represents the core of the feature-domain dereverberation algorithm. Introducing a simplified notation by the mappings $\mathbf{s}(n) \rightarrow \mathbf{s}_0$, $\check{\mathbf{s}}_{ij}(n-m) \rightarrow \check{\mathbf{s}}_m$, $\mathbf{x}(n) \rightarrow \mathbf{x}$, $\mathbf{h}(m,n) \rightarrow \mathbf{h}_m$, the constraint (4) can be written as

$$\mathbf{x} = \underline{\mathbf{h}_0} \odot \underline{\mathbf{s}_0} + \sum_{m=1}^{M-1} \underline{\mathbf{h}_m} \odot \overline{\check{\mathbf{s}}_m} \ , \tag{5}$$

where the $\underline{\textit{underlined vectors}}$ are Gaussian random vectors with diagonal covariance matrix and the $\overline{\textit{overlined vectors}}$ are known from previous steps of the algorithm.

The constraint is linearized by approximating the non-Gaussian random vector $\tilde{\mathbf{x}}_0 = \mathbf{h}_0 \odot \mathbf{s}_0$ with a Gaussian random vector $\mathbf{x}_0$ exhibiting the same mean and variance as $\tilde{\mathbf{x}}_0$. Thus we can express the constraint as

$$\mathbf{x} = \underline{\mathbf{x}_0} + \sum_{m=1}^{M-1} \underline{\mathbf{h}_m} \odot \overline{\check{\mathbf{s}}_m} \ . \tag{6}$$

A two-step solution to the inner optimization problem is derived in [1]. In the first step, the optimum vectors $\mathbf{x}_0, \mathbf{h}_m \ \forall m = 1 \ldots M-1$ are determined by solving (3) s. t. (6). The corresponding closed-form solution is given in [1]. In the second step, the optimum vectors $\mathbf{h}_0$ and $\mathbf{s}_0$ are calculated by maximizing

$$f_\lambda(j,\mathbf{s}_0) \cdot f_\eta(\mathbf{h}_0) \quad \text{s. t.} \quad \overline{\mathbf{x}_0} = \underline{\mathbf{h}_0} \odot \underline{\mathbf{s}_0} \ . \tag{7}$$

Applying the method of Lagrange multipliers to this problem yields the fourth-order equation

$$\sigma^2_{\mathbf{s}_0} \odot \mathbf{h}_0^4 - m_{\mathbf{h}_0} \odot \sigma^2_{\mathbf{s}_0} \odot \mathbf{h}_0^3 + m_{\mathbf{s}_0} \odot \sigma^2_{\mathbf{h}_0} \odot \mathbf{x}_0 \odot \mathbf{h}_0 - \mathbf{x}_0^2 \odot \sigma^2_{\mathbf{h}_0} = 0 \tag{8}$$

to be fulfilled by the optimum vectors. Therefore, numerical methods are employed in [1] to find $\mathbf{h}_0$ and $\mathbf{s}_0$. All operations in (8) have to be performed element-wise, and $m_{\mathbf{h}_0}$, $\sigma^2_{\mathbf{h}_0}$, $m_{\mathbf{s}_0}$, and $\sigma^2_{\mathbf{s}_0}$ denote the mean and the variance vectors of $\mathbf{h}_0$ and $\mathbf{s}_0$, respectively.

3. SIMPLIFIED DECODING APPROACH

In this section, two simplifications for the inner optimization problem are proposed to reduce the computational complexity of the decoding algorithm used in the generic REMOS concept. The first simplification reduces the number of vectors calculated in the first step of the inner optimization. The second simplification avoids the solution of the fourth-order equation in the second step of the inner optimization by using the mean of $\mathbf{h}_0$ as an estimate for the optimum vector $\mathbf{h}_0$.

3.1. Simplification of the first step

In the first step of the inner optimization used in [1], the optimum vectors $\mathbf{x}_0, \mathbf{h}_m$ are determined for all $m = 1 \dots M-1$ by maximizing (3) s. t. (6). Since only the vector $\mathbf{x}_0$ and the contribution of the reverberation model to the Viterbi score is needed for the subsequent steps of the algorithm, the problem can be simplified by capturing the reverberation exceeding the current frame with one random vector $\underline{\mathbf{x}_R}$ given by

$$\underline{\mathbf{x}_R} = \sum_{m=1}^{M-1} \underline{\mathbf{h}_m} \odot \overline{\check{\mathbf{s}}_m} \, . \tag{9}$$

Since $\underline{\mathbf{x}_R}$ is a weighted sum of Gaussian random vectors, $\underline{\mathbf{x}_R}$ is also Gaussian with mean and variance vectors given by

$$\begin{aligned} m_{\mathbf{x}_R} &= \sum_{m=1}^{M-1} m_{\mathbf{h}_m} \odot \overline{\check{\mathbf{s}}_m} \, , \\ \sigma^2_{\mathbf{x}_R} &= \sum_{m=1}^{M-1} \sigma^2_{\mathbf{h}_m} \odot \overline{\check{\mathbf{s}}^2_m} \, , \end{aligned}$$

where $m_{\mathbf{h}_m}$ and $\sigma^2_{\mathbf{h}_m}$ are the mean and variance vectors of $\mathbf{h}_m$. Therefore, the first step of the inner optimization reduces to determining $\mathbf{x}_0$ and $\mathbf{x}_R$ by solving the following problem

$$\max_{\mathbf{x}_0, \mathbf{x}_R} \{ \, f_{\mathbf{x_0}}(\mathbf{x}_0) \cdot f_{\mathbf{x}_R}(\mathbf{x}_R) \} \tag{10}$$

$$\text{s. t.} \quad \mathbf{x} = \underline{\mathbf{x}_0} + \underline{\mathbf{x}_R} \, . \tag{11}$$

Applying the method of Lagrange multipliers, we obtain

$$\begin{aligned} \mathbf{x}_0 &= \frac{\sigma^2_{\mathbf{x}_R}}{\sigma^2_{\mathbf{x}_0} + \sigma^2_{\mathbf{x}_R}} \odot m_{\mathbf{x}_0} + \frac{\sigma^2_{\mathbf{x}_0}}{\sigma^2_{\mathbf{x}_0} + \sigma^2_{\mathbf{x}_R}} \odot (\mathbf{x} - m_{\mathbf{x}_R}) \, , \\ \mathbf{x}_R &= \frac{\sigma^2_{\mathbf{x}_0}}{\sigma^2_{\mathbf{x}_0} + \sigma^2_{\mathbf{x}_R}} \odot m_{\mathbf{x}_R} + \frac{\sigma^2_{\mathbf{x}_R}}{\sigma^2_{\mathbf{x}_0} + \sigma^2_{\mathbf{x}_R}} \odot (\mathbf{x} - m_{\mathbf{x}_0}) \, . \end{aligned}$$

Both vectors $\mathbf{x}_0, \mathbf{x}_R$ are used for the calculation of the Viterbi score, and $\mathbf{x}_0$ is also used in the second step of the inner optimization.

3.2. Simplification of the second step

To avoid the solution of the fourth-order equation (8), we replace the optimum vector $\mathbf{h}_0$ with the mean vector $m_{\mathbf{h}_0}$, and by solving the constraint in (7) for $\mathbf{s}_0$, we obtain

$$\mathbf{s}_0 = \frac{\overline{\mathbf{x}_0}}{m_{\mathbf{h_0}}} \, . \tag{12}$$

	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.

The estimated clean-speech vector $\mathbf{s}_0$ is used in the calculation of the Viterbi score. Furthermore it is used as the basis to calculate the most likely clean speech estimate $\hat{\mathbf{s}}_j(n)$ for the current state j and the current frame n which is stored in the matrix of clean speech estimates (see [12]).

4. EXPERIMENTS

Experiments with the same connected-digit recognition task as used in [1] are carried out to analyze the performance and the computational savings of the simplified algorithm.

4.1. Experimental setup

The experimental setup is identical to that of [1]. Therefore, only the most important facts are recalled here. The REMOS-based recognizer is implemented by extending the decoding routines of HTK [13]. 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 get 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. A strict separation of training data (speech and RIRs) from the test data is maintained in all experiments. 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 the test.

4.2. Experimental results

Table 2 compares the word accuracies and the computational complexity of conventional HMM-based recognizers to that of the REMOS concept with different decoding methods. As the experiments are based on melspec features and single Gaussian densities, the recognition rates are not comparable to those of state-of-the-art recognizers using MFCCs and mixtures of Gaussians. Simplifying the first step of the inner optimization according to Section 3.1 reduces the computational complexity of the REMOS concept to 88 %, 61 %, and 53 %

	clean		Room					
	data		A		B		C	
	acc.	RTF	acc.	RTF	acc.	RTF	acc.	RTF
(I) conventional HMM, clean training	82.0 %	< 0.1	51.5 %	< 0.1	13.4 %	< 0.1	25.9 %	< 0.1
(II) conventional HMM, reverberant training	-	-	66.8 %	< 0.1	54.6 %	< 0.1	46.0 %	< 0.1
(III) original REMOS [1]	-	-	77.6 %	8.7	71.6 %	12.7	67.6 %	15.3
(IV) REMOS, 1st step simplified (see Section 3.1)	-	-	76.2 %	7.1	72.4 %	7.8	68.2 %	8.2
(V) REMOS, 1st and 2nd step simplified (see Section 3.1 and 3.2)	-	-	74.3 %	1.7	70.8 %	2.3	66.0 %	2.9

Table 2. Word accuracies (acc.) and Real-Time Factors (RTF) measured on an AMD Opteron processor with a clock rate of 1.6 GHz for the conventional HMM-based recognizer trained on clean (I) and reverberant speech (II), the REMOS concept with the original decoding method proposed in [1] (III), the simplified decoding applied to the first step of the inner optimization (IV), and to the first and second step of the inner optimization (V).

of the original complexity for the rooms A, B, and C, respectively. The computational savings increase with the length M of the reverberation model because the number of vectors $\mathbf{h}_m$ calculated by the original decoding algorithm increases with M. The recognition rates obtained by the simplified approach are equivalent to those of the original approach. If both the first step and the second step (see Section 3.2) of the inner optimization are simplified, the computational complexity of the REMOS concept is reduced to less than 20 % of the original complexity for all three rooms. With a real-time factor of 1.7, real-time capability is almost achieved for room A. At the same time, the recognition accuracy is only slightly reduced. That is, the recognition rates are still significantly higher than those achieved with conventional HMM-based recognizers even if their acoustic models are trained on matched reverberant data.

Because of the enormous computational reductions for the solution of the inner optimization problem, the major remaining complexity of the decoding algorithm lies in the backtracking performed in each iteration to select the clean-speech estimates $\check{\mathbf{s}}_{ij}(n-m)$ for the previous frames. Therefore, further complexity reductions can be expected from optimizing the backtracking routines.

5. SUMMARY AND CONCLUSIONS

Two simplifications of the decoding algorithm employed by the REMOS concept to perform feature-domain dereverberation for robust distant-talking ASR have been proposed in this contribution. A reduction in computational complexity to less than 20 % of the original complexity is achieved by these simplifications so that the REMOS concept is very close to real-time capability for the task of connected digit recognition. The recognition accuracy is only slightly reduced. One of the simplifications consists of capturing the reverberation exceeding the current frame in one random vector. Therefore, the number of independent variables for the inner optimization is significantly reduced. This reduction appears to be very attractive for implementing the REMOS concept for more powerful speech features, like MFCCs, which requires numerical optimization routines.

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] 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.

[3] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[4] B. Yegnanarayana and P. Satyanarayana Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 267–281, May 2000.

[5] N. D. Gaubitch, P. A. Naylor, and D. B. Ward, "On the use of linear prediction for dereverberation of speech," *Proc. IEEE Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, pp. 99–102, September 2003.

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

[7] H. Buchner, R. Aichner, and W. Kellermann, "TRINICON: A versatile framework for multichannel blind signal processing," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. III, pp. 889–892, May 2004.

[8] T. Hikichi, M. Delcroix, and M. Miyoshi, "Blind dereverberation based on estimates of signal transmission channels without precise information of channel order," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 1069–1072, March 2005.

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

[10] K. Kinoshita, T. Nakatani, and M. Miyoshi, "Spectral subtraction steered by multi-step forward linear prediction for single channel speech dereverberation," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 817–820, May 2006.

[11] B.-H. Juang, T. Nakatani, K. Kinoshita, and M. Miyoshi, "Joint source-channel modeling and estimation for speech dereverberation," *Proc. ISCAS 2007*, pp. 2990–2993, 2007.

[12] A. Sehr and W. Kellermann, "Towards robust distant-talking automatic speech recognition in reverberant environments," in *Topics in Speech and Audio Processing in Adverse Environments*, E. Hänsler and G. Schmidt, Eds. Springer, Berlin, to appear.

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