14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

HANDS-FREE SPEECH RECOGNITION USING A REVERBERATION MODEL IN THE FEATURE DOMAIN

Armin Sehr, Marcus Zeller, and Walter Kellermann

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

ABSTRACT

A novel approach for robust hands-free speech recognition in highly reverberant environments is proposed. Unlike conventional HMM-based concepts, it implicitly accounts for the statistical dependence of successive feature vectors due to the reverberation. This property is attained by a combined acoustic model consisting of a conventional HMM, modeling the clean speech, and a reverberation model. Since the HMM is independent of the acoustic environment, it needs to be trained only once using the usual Baum-Welch re-estimation procedure. The training of the reverberation model is based on a set of room impulse responses for the corresponding acoustic environment and involves only a negligible computational effort. Thus, the recognizer can be adapted to new environments with moderate effort. In a simulation of an isolated digit recognition task in a highly reverberant room, the proposed method achieves a 60 % *reduction of the word error rate compared to a conventional HMM trained on reverberant speech, at the cost of an increased decoding complexity.*

1. INTRODUCTION

Automatic speech recognition (ASR) is the key to numerous applications like natural human-machine interfaces, dictation systems, electronic translators and automatic information desks. To further increase the acceptance of these applications, it is desirable that the user can move freely while communicating to the system without the need of wearing a headset or any other kind of close-talking microphone.

Since the distance between speaker and microphone in such a hands-free scenario usually is in the range of one to several meters, there are two kinds of distortions that hamper ASR. Besides the desired signal, the microphone picks up reverberation of the desired signal and unwanted additive signals like background noise signals or interfering speakers. While significant progress has already been reported within the last decade regarding the robustness of ASR to additive distortions, ASR for highly reverberant environments has only recently attracted increasing attention.

For reliable speech recognition in reverberant environments, two classes of approaches are known. Either the speech signal picked up by the microphone is dereverberated prior to speech recognition, or the recognizer itself is made robust to reverberation. Both ways are currently intensively investigated.

Blind dereverberation of the speech signal is an extremely challenging task, since neither the impulse response of the acoustic path between speaker and microphone nor the speaker's signal are available. As this paper focuses on the realization of a robust recognizer, we refer to a review article [1] and to several promising methods [2, 3, 4] for a further discussion on blind dereverberation.

The most straightforward approach of obtaining an ASR system capable of working in reverberant environments is to train a conventional HMM-based recognizer using data recorded in the very enclosure where the recognizer will be deployed. To reduce the enormous effort implied in collecting a complete set of training data for each new environment of operation, artificial reverberation of clean training data has been suggested [5, 6] and has been shown to yield a noticeable improvement.

While the usual model adaptation techniques, which have been successfully applied in noisy environments, are not suitable for reverberation significantly exceeding the frame length of the recognizer, Raut et al. [7] suggest a model adaptation approach designed particularly for long reverberation. Here, the linear means of a split-state HMM are adjusted taking into account the linear means of the preceding states. Thus, the amount of necessary training data and the computational complexity is considerably reduced compared to reverberant training, and a significant improvement in recognition rate compared to a HMM trained on clean speech is reported in [7]. However, both reverberant training and model adaptation techniques suffer from the underlying assumption of any HMM-based system, namely that the current output vector depends only on the current state. This assumption prevents conventional HMMs from appropriately modeling reverberation.

In this paper, we propose a novel approach for robust speech recognition in reverberant environments, where the dependence of the current feature vector on previous vectors is implicitly accounted for by a combined acoustic model. The combined model consists of a conventional HMM, modeling the clean speech, and a reverberation model. Since the HMM is independent of the acoustic environment, it needs to be trained only once using the usual Baum-Welch re-estimation procedure. The training of the reverberation model is based on a set of room impulse responses for the corresponding acoustic environment and involves only a negligible computational effort. In this way, the recognizer can be adapted to new environments with moderate effort.

The paper is organized as follows: In the following section, the proposed approach is explained in detail. Simulations of an isolated digit recognition task, described in Section 3, show the effectiveness of the new recognizer. In Section 4, conclusions are drawn and important topics for future work are outlined.

2. THE PROPOSED APPROACH

We introduce the combined acoustic model from the perspective of feature production. Before deriving a solution for the decoding of the combined model, detailed descriptions of the reverberation model and the convolution in the feature domain, which is the basis for the combination of the two models, are given.

2.1 Feature production model

We assume that the sequence $\mathbf{X}$ of reverberant speech feature vectors $\mathbf{x}(n)$ is produced by a combination of an HMM λ describing the clean speech, and a reverberation model η as illustrated in Figure 1. This model for the production of reverberant feature vectors can be applied to any kind of speech features as long as an appropriate relation between the sequence $\mathbf{S}$ of output feature vectors $\mathbf{s}(n)$ of the clean speech model, the sequence $\mathbf{H}$ of the reverberation model output matrices $\mathbf{H}(n)$ and the sequence $\mathbf{X}$ of reverberant speech feature vectors $\mathbf{x}(n)$ can be formulated. In this paper, we apply this model to mel-frequency spectral coefficients (melspec coefficients), which

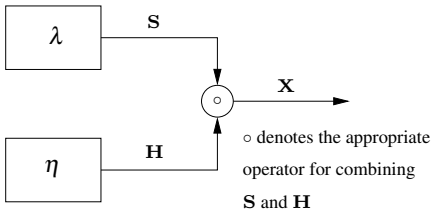

Figure 1: Proposed feature production model.

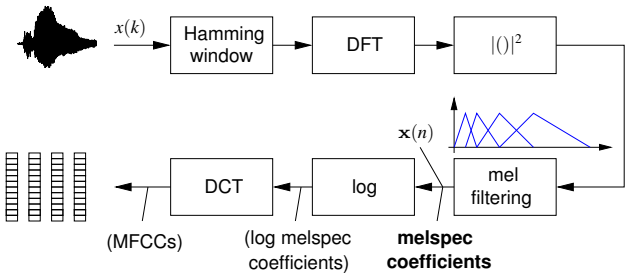

Figure 2: Calculation of melspec coefficients as a preliminary stage of MFCCs.

are the basis for the calculation of the well-known mel-frequency cepstral coefficients (MFCC) as depicted in Figure 2.

The melspec features are chosen because they allow the formulation of a very simple approximative relation between $\mathbf{S}$, $\mathbf{H}$ and $\mathbf{X}$. The reverberant speech feature sequence $\mathbf{X}$ results from the convolution of the clean speech feature sequence $\mathbf{S}$ originating from the HMM and the sequence $\mathbf{H}$ of realizations of the reverberation model.

2.2 Convolution in the melspec domain

The convolution in the melspec domain is performed according to

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

where $\odot$ denotes element-wise multiplication, $\mathbf{s}(n)$ and $\mathbf{x}(n)$ are single feature vectors at frame index n of clean and reverberant speech, respectively, and the vector $\mathbf{h}(m,n)$ is a realization of the reverberation model for frame delay m and frame index n. N and M are the lengths of the sequence $\mathbf{S}$ and the number of columns of the matrix $\mathbf{H}(n)$, respectively.

The true reverberant speech feature sequence $\mathbf{X}_t$, which is the actual input to the ASR system, however, is calculated from the reverberant speech signal $x(k)$ as illustrated in Figure 2. The signal $x(k)$ results from the linear convolution of the clean speech signal $s(k)$ and the impulse response $h(k)$ of the acoustic path between speaker and microphone. Compared to the true sequence $\mathbf{X}_t$, the computation of $\mathbf{X}$ by the convolution in the feature domain includes the following approximations:

- The constraint which had to be applied to realize an exact linear convolution by the overlap-save method [8] is neglected.
- Due to the squared magnitude operation in the feature extraction, the phase is ignored.
- Because of the mel-filtering, the frequency resolution is reduced.
- Since the order of convolution and feature extraction is reversed, the squared magnitude of a sum in the computation of $\mathbf{X}_t$ is replaced by a sum of squared magnitudes in (1).

Despite these approximations, the convolution in the melspec domain is still very closely related to the true reverberant speech features as can be verified in Figure 3 by comparing the clean feature sequence $\mathbf{S}$, the true reverberant feature sequence $\mathbf{X}_t$ and the reverberant feature sequence $\mathbf{X}$ resulting from melspec convolution for room B (see Section 3.1).

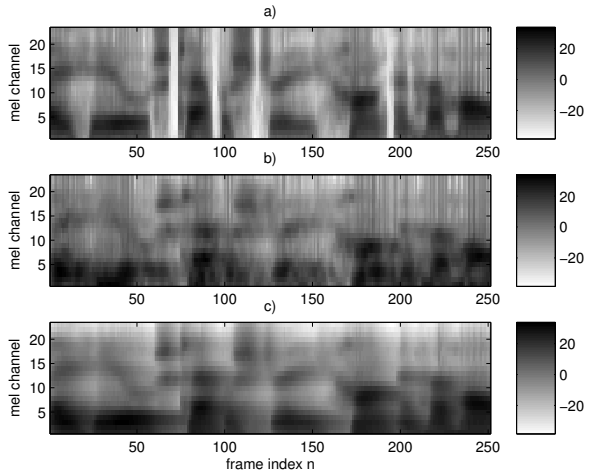

Figure 3: Comparison of a) clean feature sequence $\mathbf{S}$, b) true reverberant feature sequence $\mathbf{X}_t$, and c) reverberant feature sequence $\mathbf{X}$ calculated by melspec convolution.

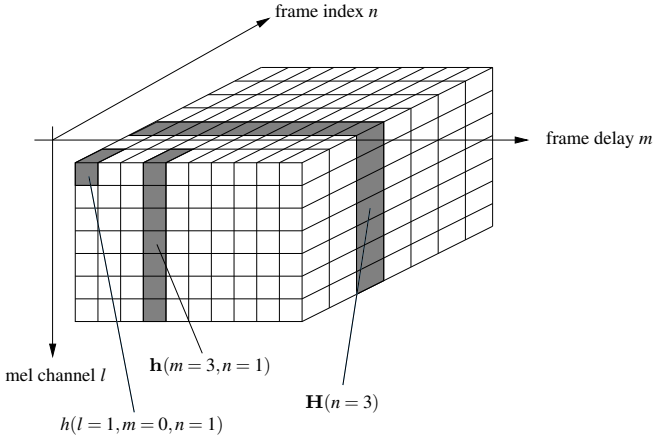

Figure 4: Realization $\mathbf{H}$ of the reverberation model η.

2.3 Reverberation model

The reverberation model η represents an independent identically distributed (iid) matrix-valued random process, where each column of the matrix corresponds to a certain delay m (in multiples of the frame shift) and each row of the matrix corresponds to a certain mel channel l. The sequence $\mathbf{H}$ of reverberation feature matrices $\mathbf{H}(n)$ is a realization of this random process as illustrated in Figure 4. For simplification, each element of the matrix is assumed to be statistically independent from all other elements and is modeled by a Gaussian density. Furthermore, the iid property of the random process implies that all elements of the random process at frame index n_1 are statistically independent from all elements of the random process at frame index n_2 as long as $n_1 \neq n_2$.

The starting point for the training of the reverberation model is a set of room impulse responses (RIRs) for different microphone and loudspeaker positions of the room where the ASR system will be applied. These RIRs can either be measured before using the recognizer, estimated by blind system identification approaches or modeled, e. g., using the image method as suggested in [9]. To train the reverberation model, the RIRs are aligned so that the direct path of all RIRs occurs at the same delay. Calculation of the melspec representation yields a matrix of melspec coefficients for each impulse response. Using these coefficients, the means and the variances of all matrix elements are estimated. Alternatively, the reverberation model can be directly estimated in the feature domain.

2.4 Decoding

So far, we introduced a novel feature production model, describing how reverberant speech features are generated given the model. For speech recognition however, the opposite task has to be solved. Given a reverberant utterance, a set of clean speech HMMs and a reverberation model, the task of the recognizer is to find the sequence of HMMs which best models the reverberant utterance in connection with the reverberation model.

For simplification, we restrict the following treatment to isolated word recognition using a single HMM for each word in the dictionary. Then the recognition process reduces to calculating the production probability for each word-level HMM and selecting the word corresponding to the most probable HMM.

Conventional HMM-based isolated word recognition uses the Viterbi algorithm or the iterative calculation of the forward probability to find the production probability for each HMM. To determine the production probabilities for the proposed combined models, an extended version of the Viterbi algorithm is introduced, which will be explained in detail in the following.

Given the reverberant speech feature sequence $\mathbf{X}_t$, a set of clean speech HMMs $\{\lambda_p\}$, where $p = 1 \ldots P$ indexes the P different word models in the dictionary, and a reverberation model η, the task of the recognizer is to find the HMM $\hat{\lambda}$ maximizing the probability that the reverberant feature sequence was produced by the combination of $\hat{\lambda}$ and η. This is achieved by maximizing the joint probability of the clean feature sequence $\mathbf{S}$, the state sequence Q of the word model λ_p and the sequence of reverberation features $\mathbf{H}$

$$\begin{aligned}\hat{\lambda} &= \underset{\lambda_p}{\operatorname{argmax}}\{P(\mathbf{S}^*, Q^*, \mathbf{H}^*|\lambda_p, \eta)\} \\ &= \underset{\lambda_p}{\operatorname{argmax}}\{\max_{\mathbf{S},Q,\mathbf{H}}\{P(\mathbf{S}, Q, \mathbf{H}|\lambda_p, \eta)\}\}\end{aligned}$$

subject to (s. t.) the constraint (1).

An approximative solution of

$$\max_{\mathbf{S},Q,\mathbf{H}}\{P(\mathbf{S}, Q, \mathbf{H}|\lambda_p, \eta)\} \quad \text{s. t. (1)}$$

is obtained by an extended version of the Viterbi algorithm for each HMM λ_p.

To simplify the notation, the subscript p is omitted in the following outline of the extended Viterbi algorithm. Defining the best-path probability of state j at frame n

$$\gamma_j(n) = \max_{\mathbf{S}(1\ldots n),Q(1\ldots n-1),\mathbf{H}(1\ldots n)} P\left(\begin{array}{l}\mathbf{s}(1)\ldots\mathbf{s}(n), \mathbf{H}(1)\ldots\mathbf{H}(n), \\ q(1)\ldots q(n-1), q(n)=j|\lambda,\eta\end{array}\right),$$

where $q(n)$ is the HMM state at frame n, and $\mathbf{S}(1\ldots n)$, $Q(1\ldots n)$ and $\mathbf{H}(1\ldots n)$ denote partial sequences from frame 1 to frame n, and the backtracking pointer $\psi_j(n)$ referring to the previous state, the extended Viterbi algorithm is given by

Init:

$$\begin{aligned}\gamma_1(1) &= \max_{\mathbf{s}(1),\mathbf{h}(0,1)}\left\{f_\lambda(1,\mathbf{s}(1))\cdot f_\eta(\mathbf{h}(0,1))\right\}, \\ &\text{s. t.} \quad \mathbf{x}(1) = \mathbf{s}(1)\cdot\mathbf{h}(0,1) \\ \gamma_j(1) &= 0 \quad \forall j = 2\ldots I, \\ \psi_j(1) &= 0 \quad \forall j = 1\ldots I.\end{aligned}$$

Recursion:

$$\begin{aligned}\gamma_j(n) &= \max_i\left\{\gamma_i(n-1)\cdot a_{ij}\cdot O_{ij}(n)\right\}, \\ \psi_j(n) &= \underset{i}{\operatorname{argmax}}\left\{\gamma_i(n-1)\cdot a_{ij}\cdot O_{ij}(n)\right\}, \\ O_{ij}(n) &= \max_{\mathbf{s}_{ij}(n),\mathbf{H}_{ij}(n)}\left\{f_\lambda(j,\mathbf{s}_{ij}(n))\cdot f_\eta(\mathbf{H}_{ij}(n))\right\}, \qquad (2) \\ &\forall \quad j = 1\ldots I, \quad n = 2\ldots N+M-1, \\ \text{s. t.} \quad \mathbf{x}(n) &= \sum_{m=0}^{M-1}\mathbf{h}_{ij}(m,n)\odot\mathbf{s}_{ij}(n-m). \qquad (3)\end{aligned}$$

Termination:

$$P(\mathbf{S}^*, Q^*, \mathbf{H}^*|\lambda,\eta) = \gamma_I(N+M-1), \quad q(N+M-1) = I.$$

Backtracking:

$$q(n) = \psi_{q(n+1)}(n+1),$$

where I is the number of states of the HMM, a_{ij} is the HMM transition probability from state i to state j, $f_\lambda(j,\cdot)$ is the HMM output density of state j and $f_\eta(\cdot)$ is the output density of the reverberation model. The subscript ij in $\mathbf{s}_{ij}(n)$, $\mathbf{h}_{ij}(m,n)$ and $\mathbf{H}_{ij}(n)$ indicates that these vectors/matrices are based on the optimum partial state sequence Q^*_{ij} given by

$$Q^*_{ij}(n-M+1\ldots n) = q^*(n-M+1),\ldots,q^*(n-2),q(n-1)=i,q(n)=j$$

from frame $n-M+1$ to frame n with current state j and previous state i.

The extension compared to the usual Viterbi algorithm consists of the inner optimization of equation (2). Applying the method of Lagrange multipliers to the inner optimization problem

$$\max_{\mathbf{s}_{ij}(n),\mathbf{H}_{ij}(n)}\{f_\lambda(j,\mathbf{s}_{ij}(n))\cdot f_\eta(\mathbf{H}_{ij}(n))\} \quad \text{s. t. (3)}$$

yields a two-step closed-form solution. For simplification, we demonstrate this solution using an example with very short reverberation. Assume the matrix of the reverberation model consists of only $M = 3$ columns corresponding to a length of the room impulse response of $1\cdot$ frame length $+\,2\cdot$ frame shift. Then the constraint equation can be written as

$$\mathbf{x}(n) = \underline{\mathbf{h}_{ij}(0,n)}\odot\underline{\mathbf{s}_{ij}(n-0)} + \underline{\mathbf{h}_{ij}(1,n)}\odot\overline{\mathbf{s}_{ij}(n-1)} + \underline{\mathbf{h}_{ij}(2,n)}\odot\overline{\mathbf{s}_{ij}(n-2)}$$

where the *underlined vectors* are unknowns following a Gaussian distribution with diagonal covariance matrix and the *overlined vectors* are known from previous steps of the algorithm. Now we approximate the generally non-Gaussian random vector $\tilde{\mathbf{x}}_i(0,n) = \underline{\mathbf{h}_{ij}(0,n)}\odot\underline{\mathbf{s}_{ij}(n-0)}$ resulting from the element-wise product of the two Gaussian random vectors $\underline{\mathbf{h}_{ij}(0,n)}$ and $\underline{\mathbf{s}_{ij}(n-0)}$ by a Gaussian random vector $\underline{\mathbf{x}_i(0,n)}$ with the same mean and variance as $\tilde{\mathbf{x}}_i(0,n)$. Assuming statistical independence of all vector elements $\underline{s_i(l,n-0)}\ \forall\, l = 1\ldots L$, we obtain for the mean vector of $\underline{\mathbf{x}_i(0,n)}$

$$m_{\mathbf{x}_i(0,n)} = m_{\mathbf{s}_{ij}(n-0)}\odot m_{\mathbf{h}_{ij}(0,n)}$$

and for the variance vector

$$\sigma^2_{\mathbf{x}_i(0,n)} = \sigma^2_{\mathbf{h}_{ij}(0,n)}\odot\sigma^2_{\mathbf{s}_{ij}(n-0)} + \sigma^2_{\mathbf{h}_{ij}(0,n)}\odot m^2_{\mathbf{s}_{ij}(n-0)} + \sigma^2_{\mathbf{s}_{ij}(n-0)}\odot m^2_{\mathbf{h}_{ij}(0,n)},$$

where the squaring operation denotes element-wise squaring. Thus, the constraint can be rewritten as

$$\mathbf{x}(n) = \underline{\mathbf{x}_i(0,n)} + \underline{\mathbf{h}_{ij}(1,n)}\odot\overline{\mathbf{s}_{ij}(n-1)} + \underline{\mathbf{h}_{ij}(2,n)}\odot\overline{\mathbf{s}_{ij}(n-2)}.$$

Introducing the simplified notation

$$\mathbf{x} = \underline{\mathbf{x}_0} + \underline{\mathbf{h}_1}\odot\overline{\mathbf{s}_1} + \underline{\mathbf{h}_2}\odot\overline{\mathbf{s}_2}, \qquad (4)$$

we obtain the following two-step solution of the constrained problem.

First step: Find $\mathbf{x}_0$, $\mathbf{h}_1$ and $\mathbf{h}_2$.
Applying the method of Lagrange multipliers to

$$\max_{\mathbf{x}_0,\mathbf{h}_1,\mathbf{h}_2}\{f_{\mathbf{x}_0}(\mathbf{x}_0)\cdot f_\eta(\mathbf{h}_1)\cdot f_\eta(\mathbf{h}_2)\} \quad \text{s. t. (4)}, \qquad (5)$$

where $f_{\mathbf{x}_0}(\mathbf{x}_0)$ is the L-dimensional Gaussian probability density of $\mathbf{x}_0$, we obtain the following solution for $\mathbf{x}_0$, $\mathbf{h}_1$ and $\mathbf{h}_2$

$$\begin{aligned}\mathbf{x}_0 &= \frac{\mathbf{s}_1^2\odot\sigma^2_{\mathbf{h}_1} + \mathbf{s}_2^2\odot\sigma^2_{\mathbf{h}_2}}{\sigma^2_{\mathbf{x}_0} + \mathbf{s}_1^2\odot\sigma^2_{\mathbf{h}_1} + \mathbf{s}_2^2\odot\sigma^2_{\mathbf{h}_2}}\odot m_{\mathbf{x}_0} \\ &+ \frac{\sigma^2_{\mathbf{x}_0}}{\sigma^2_{\mathbf{x}_0} + \mathbf{s}_1^2\odot\sigma^2_{\mathbf{h}_1} + \mathbf{s}_2^2\odot\sigma^2_{\mathbf{h}_2}}\odot\left(\mathbf{x} - \mathbf{s}_1\odot m_{\mathbf{h}_1} - \mathbf{s}_2\odot m_{\mathbf{h}_2}\right),\end{aligned}$$

$$\mathbf{h}_1 = \frac{\sigma^2_{\mathbf{x}_0} + \mathbf{s}_2^2 \odot \sigma^2_{\mathbf{h}_2}}{\sigma^2_{\mathbf{x}_0} + \mathbf{s}_1^2 \odot \sigma^2_{\mathbf{h}_1} + \mathbf{s}_2^2 \odot \sigma^2_{\mathbf{h}_2}} \odot m_{\mathbf{h}_1} + \frac{\mathbf{s}_1^2 \odot \sigma^2_{\mathbf{h}_1}}{\sigma^2_{\mathbf{x}_0} + \mathbf{s}_1^2 \odot \sigma^2_{\mathbf{h}_1} + \mathbf{s}_2^2 \odot \sigma^2_{\mathbf{h}_2}} \odot \frac{1}{\mathbf{s}_1} \odot \left(\mathbf{x} - m_{\mathbf{x}_0} - \mathbf{s}_2 \odot m_{\mathbf{h}_2}\right),$$

$$\mathbf{h}_2 = \frac{\sigma^2_{\mathbf{x}_0} + \mathbf{s}_1^2 \odot \sigma^2_{\mathbf{h}_1}}{\sigma^2_{\mathbf{x}_0} + \mathbf{s}_1^2 \odot \sigma^2_{\mathbf{h}_1} + \mathbf{s}_2^2 \odot \sigma^2_{\mathbf{h}_2}} \odot m_{\mathbf{h}_2} + \frac{\mathbf{s}_2^2 \odot \sigma^2_{\mathbf{h}_2}}{\sigma^2_{\mathbf{x}_0} + \mathbf{s}_1^2 \odot \sigma^2_{\mathbf{h}_1} + \mathbf{s}_2^2 \odot \sigma^2_{\mathbf{h}_2}} \odot \frac{1}{\mathbf{s}_2} \odot \left(\mathbf{x} - m_{\mathbf{x}_0} - \mathbf{s}_1 \odot m_{\mathbf{h}_1}\right),$$

where the squaring and the division operation denote element-wise squaring and element-wise division, respectively, and $m_{\mathbf{h}_1}$ and $\sigma^2_{\mathbf{h}_1}$ denote the mean and the variance vector of $\mathbf{h_1}$, respectively, and likewise for the other variables. Note that these solutions have an intuitive interpretation. For example, the solution for $\mathbf{h}_1$ is a weighted sum of the mean vector $m_{\mathbf{h}_1}$ and a second term related to the constraint. If $\sigma^2_{\mathbf{h}_1}$ is small compared to the other variances, the solution is dominated by $\mathbf{m_{h_1}}$. If $\sigma^2_{\mathbf{h}_1}$ is large compared to the other variances, the solution is dominated by the second term.

Second step: Find $\mathbf{h}_0$ and $\mathbf{s}_0$ given $\mathbf{x}_0$.
Applying the method of Lagrange multipliers to

$$\max_{\mathbf{s}_0, \mathbf{h}_0} \{ f_\lambda(j, \mathbf{s}_0) \cdot f_\eta(\mathbf{h}_0) \}$$

subject to the constraint

$$\overline{\mathbf{x}_0} = \underline{\mathbf{h}_0} \odot \underline{\mathbf{s}_0},$$

we obtain the following fourth-order equation to be fulfilled by the desired vector $\mathbf{h}_0$

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

where the exponents denote element-wise powers. It can be shown, that this equation has a pair of complex conjugate solutions, one real-valued positive and one real-valued negative solution. As only the real-valued positive solution achieves the maximization of the desired probability, we obtain exactly one vector $\mathbf{h}_0$ and thus exactly one vector $\mathbf{s}_0$.

Generalizing this two-step solution of the inner optimization problem to arbitrary lengths M of the reverberation model is straightforward and the corresponding general solutions for the first step are given in Appendix A.

Note that the decoding of the combined acoustic model introduced in this paper exhibits some similarities to the HMM decomposition approach proposed in [10] for additive noise. Indeed, our approach can be considered as a generalization of the HMM decomposition for a convolutive combination of the model outputs if the reverberation model is considered as a one-state HMM with matrix-valued output. However, there is a decisive difference in the evaluation of the output density of the combined model. The HMM decomposition approach proposes to integrate over all possible combinations of the outputs of the individual models to calculate the output probability of the combined feature vector. We propose to search for the most likely combination to calculate the probability of the combined feature vector. While both approaches are feasible for simple combinations like addition, the method proposed here provides significant computational savings for more complex combinations like convolution.

3. SIMULATIONS

To analyze the effectiveness of the proposed approach, simulations of an isolated digit recognition task using melspec features are carried out. The performance of the proposed approach is compared to that of conventional HMM-based recognizers (using the same melspec features) trained on clean and reverberant speech, respectively.

3.1 Experimental Setup

For the experiments, HTK [11] is used. The functionality of HTK is extended so that the proposed algorithm can be simulated using HTK. To calculate the feature vectors from the speech signal sampled at 20 kHz, the signal is decomposed into frames of length 25 ms with a frame shift of 10 ms. After applying a 1st-order preemphasis (coefficient 0.97) and a Hamming window, the frames are transformed to the frequency domain using a 512-point DFT. 24 melspec coefficients are calculated from the DFT coefficients. Only static features and no Δ and $\Delta\Delta$ coefficients are used.

For the training, 4579 connected digit utterances corresponding to 1.5 hours of speech from the TI digits [12] training data are used. The speech signals are normalized so that the average power of each digit is equal across all digit strings. For the training with reverberant speech, the data are convolved with measured room impulse responses. Impulse responses from two different rooms are used. Room A is a lab environment with a reverberation time of $T_{60} = 300\,\text{ms}$. Room B is a studio environment with a reverberation time of $T_{60} = 700\,\text{ms}$.

A 16-state left-to-right model without skips over states is trained for each of the 11 digits ('0'-'9' and 'oh'). Additionally, a three-state silence model with a backward skip from state 3 to state 1 is trained. The output densities are single Gaussians with diagonal covariance matrix. All HMMs are trained in the following way: First, single Gaussian MFCC-based HMMs are trained by 10 iterations of Baum-Welch re-estimation. Then the melspec HMMs are obtained from the MFCC HMMs by single pass retraining [13]. For the conventional HMM-based clean recognizer and for the proposed approach, identical sets of HMMs are used. The HMM sets of the conventional reverberant recognizers differ only with respect to the training data. Two distinct sets of reverberant HMMs are trained for room A and room B using data reverberated with RIRs measured in the corresponding rooms.

For the recognition, the silence model is appended to each of the 11 digit models, so that the decoding is performed on 11 concatenated models. As test-data, 2439 single digits extracted from the test utterances of the TI digits corpus are used. They are normalized in the same way as the training data, so that each digit has the same average power. The feature sequences of the clean test data are calculated directly on the extracted and normalized digits. To obtain the reverberant feature sequences, the normalized clean test signals are convolved with room impulse responses from room A and room B, respectively, before they are passed to the feature extraction unit.

To train the reverberation model η_A for room A with length $M_A = 20$, 36 impulse responses measured in room A with different loudspeaker and microphone positions with constant distance of 2.00 m are used. For the training of η_B with length $M_B = 50$, 18 impulse responses measured in room B with a loudspeaker microphone distance of 4.12 m are used. For the artificial reverberation of training data and for the training of the reverberation models, RIRs different from the impulse responses used to generate the test data (measured in the same room but at different microphone positions) are used in order to maintain a strict separation of training and test data.

3.2 Experimental Results

Table 1 compares the word error rates (WER) of the conventional HMM-based recognizers to that of the proposed approach for the isolated digit recognition task described above. The low recognition rate for the clean conventional HMM-based recognizer with clean test data results from using melspec features in connection with single Gaussian output densities providing only a coarse approximation of the true densities. While the WER increase in room A compared to clean speech is more than 22 % and more than 17 % for the conventional systems trained on clean or reverberant speech, respectively, the error rate of the proposed approach only increases by less than 2 %. The advantage of the proposed approach becomes even more dominant for the more reverberant environment

Test	HMM-based clean training	HMM-based rev. training	proposed
clean data	14.79 %	-	-
rev. data - room A	37.06 %	32.23 %	16.36 %
rev. data - room B	61.42 %	55.31 %	23.45 %

Table 1: Comparison of word error rates of a conventional HMM-based recognizer and of the proposed algorithm.

of room B. Here, the WER increases compared to the clean data performance of the conventional HMM-based recognizers are about 47 % and 41 % for clean and reverberant training, respectively, while the WER increase of the proposed approach is only about 9 %. These results confirm that the proposed approach achieves much better recognition performance in reverberant environments than conventional ASR systems, even if the latter are trained on reverberant data. However, the decoding complexity increases by a multiplicative factor which is proportional to the length M of the reverberation model. In our current implementation, this factor is in the range of one thousand.

4. CONCLUSIONS AND FUTURE WORK

We proposed a novel method tailored to recognize speech in reverberant environments which is based on a combination of an HMM and a reverberation model. The reverberant speech feature vectors are assumed to result from a feature-domain convolution of the HMM output and the output of the reverberation model. For speech recognition, an extended version of the well-known Viterbi algorithm is used to decode the unknown utterances. Simulations of isolated digit recognition in reverberant environments have shown that, at the cost of an increased decoding complexity, the proposed algorithm significantly improves the recognition rate compared to conventional HMM-based recognizers trained on reverberant speech. Future work will include generalization of the proposed approach to connected word recognition and continuous speech recognition and implementation of the method for more powerful speech features like mel-frequency cepstral coefficients.

A. GENERAL SOLUTIONS OF THE INNER OPTIMIZATION PROBLEM

If we apply the method of Lagrange multipliers to equation (5) for an arbitrary length M of the reverberation model, we get the following general solutions for the first step of the inner optimization problem

$$\begin{aligned}\mathbf{x}_0 &= \frac{\sum_{m=1}^{M-1}\mathbf{s}_m^2\odot\sigma_{\mathbf{h}_m}^2}{\sigma_{\mathbf{x}_0}^2+\sum_{m=1}^{M-1}\mathbf{s}_m^2\odot\sigma_{\mathbf{h}_m}^2}\odot m_{\mathbf{x}_0}\\ &+ \frac{\sigma_{\mathbf{x}_0}^2}{\sigma_{\mathbf{x}_0}^2+\sum_{m=1}^{M-1}\mathbf{s}_m^2\odot\sigma_{\mathbf{h}_m}^2}\odot\left(\mathbf{x}-\sum_{m=1}^{M-1}\mathbf{s}_m\odot m_{\mathbf{h}_m}\right),\end{aligned}$$

$$\begin{aligned}\mathbf{h}_{m'} &= \frac{\sigma_{\mathbf{x}_0}^2+\sum_{\substack{m=1\\m\neq m'}}^{M-1}\mathbf{s}_m^2\odot\sigma_{\mathbf{h}_m}^2}{\sigma_{\mathbf{x}_0}^2+\sum_{m=1}^{M-1}\mathbf{s}_m^2\odot\sigma_{\mathbf{h}_m}^2}\odot m_{\mathbf{h}_{m'}}\\ &+ \frac{\mathbf{s}_{m'}^2\odot\sigma_{\mathbf{h}_{m'}}^2}{\sigma_{\mathbf{x}_0}^2+\sum_{m=1}^{M-1}\mathbf{s}_m^2\odot\sigma_{\mathbf{h}_m}^2}\odot\frac{1}{\mathbf{s}_{m'}}\odot\left(x-m_{\mathbf{x}_0}-\sum_{\substack{m=1\\m\neq m'}}^{M-1}\mathbf{s}_m\odot m_{\mathbf{h}_m}\right),\end{aligned}$$

where the squaring and the division operation denote element-wise squaring and element-wise division, respectively.

REFERENCES

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

[2] T. Nakatani, M. Miyoshi, and K. Kinoshita, "Implementation and effects of single channel dereverberation based on the harmonic structure of speech," *Proc. IEEE Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, pp. 91–94, September 2003.

[3] 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.

[4] 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, 2005.

[5] 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.

[6] T. Haderlein, E. Nöth, W. Herbordt, W. Kellermann, and H. Niemann, "Using artificially reverberated training data in distant-talking ASR," *Proc. International Conference on Text, Speech and Dialogue*, pp. 226–229, 2005.

[7] C. K. Raut, T. Nishimoto, and S. Sagayama, "Model adaptation by state splitting of HMM for long reverberation," *International Conference on Spoken Language Processing (ICSLP 2005)*, pp. 277–280, September 2005.

[8] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications*, Prentice Hall, Upper Saddle River, NJ, USA, 1996.

[9] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, April 1979.

[10] A. P. Varga and R. K. Moore, "Hidden markov model decomposition of speech and noise," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 845–848, 1990.

[11] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.2)*, Cambridge University Engineering Department, Cambridge, UK, 2002.

[12] 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.

[13] 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.