FRAME-WISE HMM ADAPTATION USING STATE-DEPENDENT REVERBERATION ESTIMATES

Armin Sehr, Roland Maas, and Walter Kellermann

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

ABSTRACT

A novel frame-wise model adaptation approach for reverberation-robust distant-talking speech recognition is proposed. It adjusts the means of static cepstral features to capture the statistics of reverberant feature vector sequences obtained from distant-talking speech recordings. The means of the HMMs are adapted during decoding using a state-dependent estimate of the late reverberation determined by joint use of a feature-domain reverberation model and optimum partial state sequences. Since the parameters of the HMMs and the reverberation model can be estimated completely independently, the approach is very flexible with respect to changing acoustic environments. Due to the frame-wise model adaptation, two limitations characterizing conventional HMMs, namely the assumptions of conditional independence and piecewise stationarity, are relaxed so that a very high modeling accuracy for reverberant FVSs can be achieved.

The paper is structured as follows: A way of describing the effect of reverberation on FVSs by a feature-domain RM is discussed in Sec. 2. Sec. 3 explains the proposed frame-wise model adaptation approach. Recognition results for a connected digit recognition task are presented in Sec. 4, and Sec. 5 concludes the paper.

1. INTRODUCTION

The dispersive effect of reverberation on the feature representation of speech used for Automatic Speech Recognition (ASR) changes the statistical properties of reverberant feature vector sequences significantly. Therefore, if HMMs trained on close-talking data are used, the mismatch between the acoustic models and the input features leads to a severe degradation of ASR performance. A promising approach for reducing this mismatch is to adapt the HMMs to the current reverberation condition either using a parametric model of reverberation as in [1–3] or using general adaptation methods, such as MLLR [4] or MAP [5].

The main limitation of the aforementioned approaches based on adjusting the parameters of conventional HMMs is the conditional independence assumption underlying the HMMs. This assumption states that the current feature vector only depends on the current state and not on the previous feature vectors. Therefore, conventional HMMs can capture the interframe-dependencies caused by reverberation only indirectly by dynamic features, like \( \Delta \) and \( \Delta \Delta \) features, even if context-dependent HMMs are used. To overcome this limitation, adaptation of the mean vectors in each frame based on first-order linear prediction is suggested in [6].

This paper proposes a novel frame-wise model adaptation approach for robust distant-talking speech recognition which adjusts the means of static Mel Frequency Cepstral Coefficients (MFCCs) to capture the statistics of reverberant feature vector sequences (FVSs). The means of the HMMs are adapted during decoding time using a state-based estimate of late reverberation determined from a feature-domain Reverberation Model (RM) and optimum partial state sequences. Since the parameters of the HMMs and the RM can be estimated completely independently, the approach is very flexible with respect to changing acoustic environments. Due to the frame-wise model adaptation, two limitations characterizing conventional HMMs, namely the assumptions of conditional independence and piecewise stationarity [7], are relaxed so that a very high modeling accuracy for reverberant FVSs can be achieved.

2. FEATURE-DOMAIN REVERBERATION MODELING

Neglecting additive interferences, the reverberant microphone signal \( x(n) \) can be described by a convolution of the clean-speech signal \( s(n) \) and the Room Impulse Response (RIR) \( h(n) \) according to

\[
x(n) = h(n) \ast s(n).
\]

This linear convolution in the time domain can be approximated by a convolution in the mel-spectral (melspec) domain according to [8]

\[
x_{\text{mel}}(k) \approx \sum_{m=0}^{M-1} h_{\text{mel}}(m) \odot s_{\text{mel}}(k-m),
\]

where \( s_{\text{mel}}(k), h_{\text{mel}}(m), \) and \( x_{\text{mel}}(k) \) are the melspec feature vectors corresponding to the clean speech, the RIR, and the reverberant speech, respectively, and \( \odot \) denotes element-wise multiplication.

To describe the reverberant FVS \( x_{\text{mel}}(k) \) using (2), the melspec representation \( h_{\text{mel}}(m) \) of the RIR has to be known. In this paper, a statistical RM \( \eta \) as proposed in [8] is used instead of the melspec representation of a single RIR. The RM captures the statistical properties of RIRs for a set of speaker and microphone positions in a certain room. It exhibits a matrix structure, where each row corresponds to a certain mel channel \( l \) and each column to a certain frame delay \( m \) as shown in Fig. 1. Each of the matrix elements is modeled by an independent identically distributed Gaussian random process in the logmelspec or MFCC domain. For simplicity, different matrix elements are assumed to be statistically independent (see [8]). Therefore, the RM of length \( M \) is completely described by its mean matrix \( \mu_H(m=0:M-1) \) and its covariance matrix \( \sigma_H^2(m=0:M-1) \), where \( m = 0 : M - 1 \) denotes all frames.
from 0 to $M - 1$. Various ways to estimate these parameters are discussed in [8].

3. FRAME-WISE MODEL ADAPTATION APPROACH

The modeling accuracy of parametric HMM adaptation approaches that adjust the HMM parameters before recognition, like [1–3], is mainly limited by the fact that the reverberation can only be described on average based on the average state occupation times. Adaptation is particularly difficult for the initial states as the reverberation in these states strongly depends on the left HMM context.

One possibility to overcome these inaccuracies is to adapt the HMMs on a frame-by-frame basis during recognition. In contrast to the conventional parametric adaptation methods, [1–3], the model adaptation is performed within the main recognition loop. Thus, the most likely partial state sequence corresponding to the current utterance can be determined so that a significantly more accurate reverberation estimate can be obtained. Since the most likely partial path to the current state implicitly provides a left context, the adaptation of the initial HMM states benefits in particular from the frame-wise adaptation scheme.

In the following, a frame-wise adaptation scheme based on the RM discussed in Sec. 2 is introduced. Since it employs the same RM and parts of the decoding procedure as the REMOS (REverberation MOdeling for Speech Recognition) concept described in [8], the proposed frame-wise adaptation approach can be considered as a variant of the REMOS framework and will be denoted REMOS-FMA, for REMOS Frame-wise Model Adaptation. The main differences to the generic REMOS approach of [8] are discussed after explaining the frame-wise adaptation algorithm.

Note that the REMOS-FMA concept exhibits some similarities to the frame-wise adaptation scheme suggested by Takiguchi in [6]. The main difference between the two concepts lies in the calculation of the reverberation. Since REMOS-FMA uses a more detailed RM than [6], it can be expected that it can capture the statistical properties of reverberant FVSs more accurately. Furthermore, the reverberation estimate of [6] is purely feature-based, that is, the same reverberation estimate is added to the means of all states. In contrast, the reverberation estimate of REMOS-FMA is calculated from the respective clean-speech models as explained below, and thus it is state-dependent. Therefore, adding different reverberation estimates to different states leads to an increased discrimination capability.

3.1. HMM Adaptation Embedded into the Viterbi Algorithm

In the following, first an overview of the adaptation embedded into the Viterbi algorithm is given, and then the algorithmic details are described. Before performing the Viterbi recursion for the current frame, the HMM parameters are adapted in two steps as shown in Fig. 2. First, the means and variances of the 0-th column of the reverberation model are added to the means and variances directly in the MFCC domain. The resulting means are then transformed to the melspec domain, where the state-dependent late reverberation estimate is added. After transforming the mean vectors back to the MFCC domain, they are used for calculating the Viterbi score.

A detailed algorithmic description is given in the following: The conventional Viterbi recursion determines the Viterbi score $\gamma_j(k)$ of the current frame $k$ and the current state $Q(k) = j$ by maximizing the product of the previous Viterbi scores and the state transition probabilities $a_{ij}$ according to

$$
\gamma_j(k) = \max_i \left\{ \gamma_i(k-1) \cdot a_{ij} \cdot f_{\mathbf{s}_{\text{cep}}}(k) | Q(k) = j \right\} .
$$

Since the output pdf $f_{\mathbf{s}_{\text{cep}}}(k) | Q(k) = j$ of the current state $j$ from the clean-speech HMM, evaluated for the reverberant observation vector $\mathbf{x}_{\text{cep}}(k)$, is independent of the previous state $i$, it can simply be multiplied with the maximum. This recursion equation is replaced in the REMOS-FMA approach by

$$
\gamma_j(k) = \max_i \left\{ \gamma_i(k-1) \cdot a_{ij} \cdot f_{\mathbf{s}_{\text{cep}}}(k) | Q(k-1) = i \right\} .
$$

As now the adapted output pdf $f_{\mathbf{x}_{\text{cep}}}(k) | Q(k) = j$, $Q(k-1) = i$ of the current state $Q(k) = j$ for the observed reverberant feature vector $\mathbf{x}_{\text{cep}}(k)$ depends on the previous state $Q(k-1) = i$, the multiplication with the output pdf has to be performed within the maximization operation.

Calculation of the Gaussian pdf $f_{\mathbf{x}_{\text{cep}}}(k) | Q(k) = j$, $Q(k-1) = i$ in the MFCC domain is described by the following algorithm:

1) Add the mean vector $\mu_{\mathbf{H}_{\text{cep}}(m=0)}$ of the 0-th column of the RM to the mean vector $\mu_{\mathbf{s}_{\text{cep}}(k)}$ of the current state from the clean-speech (unadapted) HMM

$$
\mu_{\mathbf{x}_{\text{cep}}(k)} = \mu_{\mathbf{s}_{\text{cep}}(k)} + \mu_{\mathbf{H}_{\text{cep}}(m=0)}
$$

2) Transform $\mu_{\mathbf{x}_{\text{cep}}(k)}$ and $\mu_{\mathbf{H}_{\text{cep}}(m)}$ to the melspec domain to obtain $\mu_{\mathbf{x}_{\text{cep}}(k)}$ and $\mu_{\mathbf{H}_{\text{cep}}(m)}$.

3) Add the reverberation estimate $\hat{\mathbf{x}}_{r,ij,mel}(k)$ to $\mu_{\mathbf{x}_{\text{cep}}(k)}$, $\mu_{\mathbf{H}_{\text{cep}}(m)}$ (adapted)

$$
\mu_{\mathbf{x}_{\text{cep}}(k)} = \mu_{\mathbf{x}_{\text{cep}}(k)} + \mathbf{\hat{x}}_{r,ij,mel}(k)
$$

4) Add the late reverberation estimate $\mathbf{s}_{ij,mel}(k)$

$$
\mathbf{s}_{ij,mel}(k) = \sum_{m=1}^{M} \mu_{\mathbf{H}_{\text{cep}}(m)} \circ \mathbf{s}_{ij,mel}(k-m) .
$$

The determination of the clean-speech estimates $\mathbf{s}_{ij,mel}(k-m)$ is described in Sec. 3.2.
3.2. Calculation of the Clean-Speech Estimates

The calculation of the late reverberation estimate in (8) requires the clean-speech estimates \( \hat{q}_{ij}(k-M+1 : k) \) corresponding to the optimum partial state sequence \( \hat{q}_{ij}(k-M+1 : k) \) from frame \( k-M+1 \) to frame \( k \) with current state \( j \) and previous state \( i \) given by

\[
\hat{q}_{ij}(k-M+1 : k) = \hat{q}_{ij}(k-M+1), \ldots, \hat{q}_{ij}(k-2), \hat{q}_{ij}(k-1) = i, \hat{q}_{ij}(k) = j
\]

as illustrated for two different previous states in Fig. 3.

The clean-speech estimate \( \hat{s}_{ij,mel}(k') \) for frame \( k' \) and state \( j' \) is simply obtained by transforming the mean vector \( \mu_{s_{ij,mel}(k')} \) of the \( j' \)-th state from the clean-speech HMM to the melspec domain. For each \( k' \) and \( j' \), this clean-speech feature estimate \( \hat{s}_{ij,mel}(k') \) is stored in a matrix of clean-speech vectors (3D tensor). Since the matrix of clean-speech vectors is filled up to column \( k-1 \) by the previous iterations before the recursion for frame \( k \) starts, the estimated clean-speech vectors \( \hat{s}_{ij,mel}(k-m) \) can be obtained from this matrix using the optimum partial path \( \hat{q}_{ij}(k-M+1 : k) \).

4. EXPERIMENTS

4.1. Experimental Setup

The REMOS-FMA concept is implemented by extending the functionality of HTK [9] with the frame-wise model adaptation approach described in Sec. 3. The simulations are performed using RIRs measured in three different rooms with characteristics given in Table 1. Note that room A is a moderately reverberant environment while room B and room C are highly reverberant environments. A set of RIRs is measured for different loudspeaker and microphone positions in each room. Each set of RIRs is split into disjoint sets, one used for training and the other used for test. Thus, a strict separation of training and testing data is achieved as detailed in [8].

The feature vectors are calculated in the following way: The speech signal, sampled at 20 kHz, is decomposed into overlapping frames of length 25 ms with a frame shift of 10 ms. After applying a first-order pre-emphasis filter (coefficient 0.97) and a Hamming window, a 512-point DFT is computed. From the DFT representation, 12 MFCCs including the 0-th coefficient are calculated.
### Table 2. Summary of recognition results.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Clean</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. clean HMM</td>
<td>97.1%</td>
<td>82.9%</td>
<td>52.4%</td>
<td>44.1%</td>
</tr>
<tr>
<td>2. reverberant HMM</td>
<td>-</td>
<td>90.8%</td>
<td>84.8%</td>
<td>75.4%</td>
</tr>
<tr>
<td>3. adapted HMM [2]</td>
<td>-</td>
<td>84.5%</td>
<td>59.0%</td>
<td>56.8%</td>
</tr>
<tr>
<td>4. adapted HMM [3]</td>
<td>-</td>
<td>87.2%</td>
<td>79.9%</td>
<td>69.1%</td>
</tr>
<tr>
<td>5. adapted HMM – MLLR</td>
<td>-</td>
<td>89.2%</td>
<td>79.9%</td>
<td>69.7%</td>
</tr>
<tr>
<td>5. REMOS-FMA (proposed)</td>
<td>-</td>
<td>94.3%</td>
<td>87.5%</td>
<td>83.2%</td>
</tr>
</tbody>
</table>

A 16-state left-to-right model without skips over states is trained for each of the 11 digits (‘0’–‘9’ and ‘oh’). The output densities are single Gaussians with diagonal covariance matrices. For the training, 4579 connected digit utterances corresponding to 1.5 hours of speech from the TI digits [10] training data are used. For the training with reverberant speech, the clean training data are convolved with randomly selected measured RIRs from the training set of the corresponding room. MLLR adaptation is performed for the means only using 44 calibration utterances of different speakers from the task environment. The adaptation is based on the data-driven regression class tree clustering method provided by HTK. As test data, 512 test utterances randomly selected from the TI digits test set are used. The clean test data are convolved with RIRs randomly selected from the test set of the corresponding rooms to obtain the reverberant test data. To train the RMs for each room, the measured RIRs from the corresponding training set are used according to the procedure described in [8].

#### 4.2. Experimental Results

Table 2 compares the word accuracies achieved with REMOS FMA in rooms A, B, and C to a number of other adaptation and training methods. The word accuracy of the HMMs trained on clean-speech (1.) decreases steeply with increasing reverberation from room A to room C, underlining the necessity for model adjustment. Using HMMs trained on matched reverberant data, the word accuracies are significantly increased compared to the clean HMMs in all rooms. The adaptation approaches of [2] and [3] lead to noticeable and consistent gains in word accuracy for all configurations compared to the clean HMMs. The performance of MLLR adaptation is slightly better than that of [3]. However, the performance of all conventional HMM adaptation approaches is significantly lower than that of the reverberantly-trained HMMs.

REMOS-FMA significantly outperforms the reverberantly-trained HMMs in all three rooms. Since for changing reverberation conditions, only the RM has to be adjusted, REMOS-FMA is also significantly more flexible than matched reverberant training, which requires a large set of training data and a complete retraining of the HMMs for each new acoustic environment. The performance gain of REMOS-FMA can be attributed to the relaxation of two HMM limitations, namely piecewise stationarity within frames and conditional independence [7]. Since the reverberation estimate $\hat{x}_{r,i,j,\text{mel}}(k)$ (see Sec. 3.1) is based on the optimum partial state sequence obtained from the reverberant observations, the output density in each state indirectly depends on these observations and thus the conditional independence assumption is alleviated. The piecewise stationarity assumption is mitigated because the reverberation estimate $\hat{x}_{r,i,j,\text{mel}}(k)$ is different for each frame so that the output density of the same state changes for each frame. For illustration of how changing output densities within states may improve modeling accuracy, consider the first state modeling the phoneme /l/ of the HMM for the digit “seven” in the logmel spec domain. Assuming that this state is modeling four successive frames in a certain utterance, the output densities for the high-frequency mel channels will be strongly influenced by the reverberation of the preceding phoneme /s/ so that the corresponding mean values have to be increased significantly in the first of the four frames. In the following frames, the reverberation of the preceding phoneme /s/ is decaying and therefore, the mean value has to be decreased compared to the first frame.

Due to the frame-wise adaptation scheme, the decoding complexity of REMOS-FMA is increased compared to that of the other approaches. While the conventional decoding approaches achieve Real-Time Factors (RTFs) below 0.1 for the given CDR task on a 1.6 GHz AMD Opteron processor, the RTF of REMOS-FMA varies between 1.8 and 3.2 depending on the length $M$ of the RM.

## 5. CONCLUSIONS

A frame-wise model adaptation approach for adjusting the means of static cepstral features to the statistics of reverberant feature vector sequences has been proposed and evaluated in this paper. The mean adaptation during decoding time based on a feature-domain reverberation model and optimum partial state sequences exploits the distant-talking observations for obtaining a reverberation estimate. Thus, two limiting assumptions of conventional HMMs, namely piecewise stationarity and conditional independence, are relaxed, and the word accuracy of reverberantly-trained HMMs is outperformed in a connected digit recognition experiment. Since the original clean-speech models can be used and the parameters of the HMMs and the reverberation model can be estimated completely independently, the approach is very flexible with respect to changing acoustic environments. This flexibility is obtained at the cost of a moderately increased decoding complexity. Future work includes extending the concept to dynamic features and Gaussian mixture densities.

## 6. REFERENCES


