A Novel Approach for Matched Reverberant Training of HMMs using Data Pairs

Armin Sehr, Christian Hofmann, Roland Maas, and Walter Kellermann

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

Abstract

For robust distant-talking speech recognition, a novel HMM training approach using data pairs is proposed. The data pairs of clean and reverberant feature vectors, also called stereo data, are used for deriving the HMM parameters of a matched-condition reverberant HMM from a well-trained clean-speech HMM in two steps. In the first step, the alignment of the frames to the states is determined from the clean data and the clean-speech HMM. This state-frame alignment (SFA) is then used in the second step to estimate the Gaussian mixture densities for each state of the reverberant HMM by applying the Expectation Maximization (EM) algorithm to the reverberant data. Thus, a more accurate temporal alignment is achieved than by standard matched condition training, and the discrimination capability of the HMMs is increased. Connected digit recognition experiments show that the proposed approach decreases the word error rate (WER) by up to 44 % while substantially reducing the training complexity. These improvements will make reverberant training attractive for a wider range of applications.

Index Terms: HMM training, reverberation, robust ASR, distant-talking speech recognition, stereo data.

1. Introduction

Due to reverberation, the acoustic environment has a major influence on the statistical properties of the feature vectors extracted from distant-talking recordings. Thus, if acoustic models trained on close-talking data are used with distant-talking recordings, a substantial mismatch between the models and the test data leads to a significant reduction of the recognition accuracy compared to clean-speech recordings as shown in [1].

A promising way to reduce this mismatch is to train the acoustic models with training data from the room where the recognizer is to be used. The large effort for collecting a sufficient amount of training data for each target environment prevents this form of matched reverberant training from being widely used in real-world applications. The collection of training data can be avoided if synthetically generated training data are used for reverberant training as suggested by [2, 3]. Here, only the Room Impulse Response (RIR) of the target room needs to be measured or estimated, and then, a set of reverberant training data is generated by convolving clean-speech training data with this RIR. Even though, the Lombard effect (see, e.g., [4]) and the time-variance of the acoustic path between speaker and microphone are not captured by the synthetically reverberated training data, experiments in [5] show that the performance of HMMs trained on such data is only slightly lower than that of HMMs trained on matched recorded data.

However, the computational effort for a complete training of the acoustic model with reverberated data is still too high for many applications. Therefore, a novel approach, tailored particularly to training HMMs with stereo data consisting of clean and reverberated feature vectors, is proposed in this paper. Since the main idea of this novel training algorithm is to optimally combine the information of the stereo data, we call it ICEWIND (Information Combining Estimation With Non-reverberant Data).

The paper is structured as follows: In Section 2, a generic description of the problem to be solved by ICEWIND is given and the notation for the following discussions is introduced. After presenting the main ideas of several conventional training approaches in Section 3, the details of ICEWIND are described in Section 4. Connected digit recognition experiments are performed in Section 5 to compare the performance of ICEWIND to several known training approaches, and conclusions are drawn in Section 6.

2. Problem description and notation

In the following, we describe a novel HMM training algorithm using ordered pairs of training data \((y_1(k), y_2(k))\) for \(k = 1 : K\), where \(1 : K\) denotes all frames from 1 to \(K\), in order to determine HMM parameters \(\lambda_{ij}\) describing the data \(y_i(k)\). The data \(y_i(k)\) only provide additional information which can be used to improve the training. Furthermore, it is assumed that well-trained HMM parameters \(\lambda_{ij}\) describing the data \(y_i(k)\) are already available. This means that the novel training algorithm derives HMM parameters \(\lambda_{ij}\) from the known parameters \(\lambda_{ij}\) given an ordered pair of training data.

In the following, we consider training with pairs of clean-speech and synthetically reverberated data \((s(k), x(k))\) as an example for using ICEWIND. However, we emphasize that ICEWIND is not restricted to this example but can be applied whenever an ordered pair of training data \((y_1(k), y_2(k))\) and a well-trained HMM \(\lambda_{ij}\) are available. In particular, it can also be used with pairs of clean-speech and noisy data as well as with recordings of close-talking and distant-talking data.

Neglecting additive interferences, the time-domain signal \(x(n)\) recorded by a distant microphone is obtained by the convolution of the clean-speech signal \(s(n)\) and the RIR \(h(n)\)

\[ x(n) = h(n) * s(n), \]

where \(n\) denotes the discrete-time sample index. The training data \(s(k)\) and \(x(k)\), where \(k\) is the frame index, are obtained by extracting feature vectors (e.g., MFCCs) from the signals \(s(n)\) and \(x(n)\), respectively.

We formulate the novel training algorithm in the following for continuous Gaussian-mixture-density HMMs. However, it can also be applied to semicontinuous HMMs and to HMMs with discrete output probabilities. The output density \(b_j(x(k)) = p(x(k) | q(k) = j)\) for the \(j\)-th state of a continuous mixture-density HMM is given as

\[ b_j(x(k)) = \sum_{m=1}^{M} \frac{\lambda_{jm}}{\sigma^2} N(x(k) | \mu_{jm}, \Sigma_{jm}), \]
where \( w_{jm}, \mu_{jm}, \text{ and } C_{jm} \) are the weight, mean vector, and covariance matrix of component \( m \) for state \( j \), respectively, and \( Q(k) \) is a random process of state indices.

3. Conventional training approaches

As the basis for explaining ICEWIND, we first briefly sketch the main ideas of Baum-Welch training, Viterbi training, and Single-Pass Retraining (SPR) for continuous mixture-density HMMs in the following. The Maximum Likelihood (ML) estimates for the HMM parameters are determined by the Baum-Welch algorithm [6] in an iterative estimation process consisting of the two following main steps:

1. Align each vector \( x(k) \) of the training sequence \( x(1 : K) \) to all states \( j \) and all mixture components \( m \) of the HMM set with the posterior probability \( P(Q(k) = j, R(k) = m|\{x(1 : K)\}, \lambda_x) \) of selecting state \( j \) and mixture weight \( m \) (soft alignment). Here, \( R(k) \) is a random process of mixture component indices.

2. Based on this alignment, estimate the transition probabilities, the mixture weights, the means, and the covariance matrices for each state and each mixture component.

The estimates obtained from steps 1 and 2 form a new parameter set \( \lambda_x \) which is used in the next iteration. The two steps are repeated either for a fixed number of iterations or until some termination condition is fulfilled (see, e.g., [7] for details).

In contrast to Baum-Welch estimation, Viterbi training [8] uses a hard alignment of each feature vector to the state with the highest posterior probability. This means that each feature vector is aligned to exactly one state. Otherwise, Viterbi training is identical to Baum-Welch training.

SPR, introduced in [9] for training HMMs on noise-corrupted speech, is a special form of Baum-Welch training that makes explicit use of stereo data and an already available clean-speech HMM \( \lambda_s \). First, SPR determines the soft alignment of the feature vectors \( s(k) \) to the states and mixture components of \( \lambda_s \). Based on this alignment, the transition probabilities and the mixture weights are re-estimated. Then, the mean vectors and the covariance matrices of the HMM \( \lambda_s \) are estimated based on this clean-speech alignment using the reverberant data \( x(k) \) in a single pass. That is, the parameters of each mixture component are estimated from the reverberant data using the frame-state-mixture alignment from the clean data.

For reverberant training with stereo data, all of these approaches have disadvantages: The Baum-Welch and the Viterbi approaches do not exploit the information of the clean-speech data as they only use the reverberant data. Even though SPR makes explicit use of the stereo data, the parameters of the mixture densities cannot be adjusted optimally to the changed statistics of reverberant data compared to clean-speech data (see [10] for a detailed analysis) in a single pass.

4. The novel training algorithm

We first sketch the main idea of ICEWIND, and then we give a brief algorithmic description of this novel training method. ICEWIND determines the temporal structure of speech by hard aligning the clean-speech training data to the states of a well-trained clean-speech HMM \( \lambda_s \) (Viterbi alignment). This is the optimum SFA that can be achieved since the clean-speech signal is the actual source of information providing the most accurate picture of the temporal signal structure. Therefore, the clean-speech SFA is fixed for the complete training procedure and thus determines a fixed set of training frames for each HMM state. Now the parameters of the Gaussian-mixture densities are determined by applying the standard EM algorithm to the predetermined training data of each state.

The ICEWIND algorithm can be summarized as follows:

1. Determine the SFA by hard-aligning \( s(1 : K) \) to \( \lambda_s \):

\[
\hat{q}(k) = \text{argmax}_j P(Q(k) = j|s(1 : K), \lambda_s)
\]

This SFA is fixed for the complete training procedure and is used for estimating the mixture-density parameters. Therefore, the state transition probabilities of the clean-speech HMM \( \lambda_s \) are simply copied to the reverberant HMM \( \lambda_r \). Note that the SFA \( \hat{q}(k) \) can be determined in advance as soon as the clean-speech data \( s(1 : K) \) and the clean-speech HMM \( \lambda_s \) are available. If HMMs for several different reverberation conditions with identical clean-speech data are trained, the SFA has to be performed only once and can then be used for all conditions.

2. Determine the parameters of the Gaussian-mixture density for each state \( j \) applying the standard EM algorithm to the reverberant data:

a) E-step: Calculate the posterior mixture probability for each frame \( k \) and each mixture component \( m \):

\[
\gamma_{jm}(k) = P(R(k) = m|s(k), \hat{q}(k) = j, \lambda_r)
\]

\[
= \frac{w_{jm} \mathcal{N}(x(k)|\mu_{jm}, C_{jm})}{\sum_{m=1}^{M} w_{jm} \mathcal{N}(x(k)|\mu_{jm}, C_{jm})}
\]

b) M-step: Estimate the mixture density parameters for each component \( m \):

\[
\tilde{w}_{jm} = \frac{1}{\sum_{k=1}^{K} \gamma_{jm}(k)}
\]

\[
\tilde{\mu}_{jm} = \frac{1}{\sum_{k=1}^{K} \gamma_{jm}(k)} \sum_{k=1}^{K} \gamma_{jm}(k) \cdot x(k)
\]

\[
\tilde{C}_{jm} = \frac{1}{\sum_{k=1}^{K} \gamma_{jm}(k)} \sum_{k=1}^{K} \gamma_{jm}(k) \cdot (x(k) - \tilde{\mu}_{jm})^T 
\]

The HMM parameters are updated with the estimates from step 2 so that a new parameter set \( \lambda_r \) is obtained, which is used for the following iteration. The steps 2 a) and 2 b) are repeated until some termination condition is fulfilled.

The proposed ICEWIND training shares some similarities to SPR since it also uses stereo data. However, there are two major differences: First, SPR does a soft alignment of the frames to the states, while ICEWIND does a hard Viterbi alignment. Second and most important, ICEWIND does the alignment of the frames to the Gaussian mixture components using the reverberant data while SPR performs this alignment on the clean data. Thus, ICEWIND can capture changes of the probability density function (pdf) shape due to reverberation more accurately than SPR as shown in Section 5.3.

5. Experimental analysis

The following experiments first analyze whether the HMMs trained by ICEWIND indeed lead to a more accurate temporal alignment also for the test data. Then, connected digit recognition experiments are performed to compare the recognition rates of HMMs trained with ICEWIND to those trained with conventional Baum-Welch estimation and SPR.

5.1. Experimental setup

The connected-digit recognition task is chosen for evaluation since it does not require a language model. Thus, the recognition rate is solely determined by the quality of the acoustic
model. As speech features, 12 MFCCs including the 0-th coefficient plus the corresponding \( \Delta \)-coefficients, extracted from the speech signal sampled at 20 kHz, are used. The experiments are performed using RIRs measured in five different rooms whose characteristics are summarized in Table 1. Note that room A is a moderately reverberant environment while rooms B to E are highly reverberant. A set of RIRs is measured for different loudspeaker and microphone positions in each room and is split into two disjoint sets, one used for training and the other used for test. The RIRs for rooms D and E are taken from the RWCP sound scene database [11]. The strict separation between test and training RIRs assures that the reverberation conditions of the test data have not been observed during training.

The TI digits corpus [12] is used both for test and training. A subset of 512 utterances randomly selected from the TI digits test set, corresponding to approximately 16 minutes of speech, is used for test. To obtain the reverberant test data, the clean data are convolved with RIRs randomly selected from the RIR test set of the corresponding rooms. By changing the RIRs for each utterance, the time-variance of the acoustic path between speaker and microphone is simulated. The decoding of all HMMs is performed with the HTK [13] tool HVITÉ.

A subset of the TI digits training set with 4579 connected digit utterances, corresponding to 1.5 hours of speech, is used for training. By convolving the clean-speech signal \( s(n) \) from the training set with RIRs randomly selected from the RIR training set of the corresponding room according to (1), the reverberant signal \( x(n) \) is obtained. From these signals, the stereo data \( (s(k), x(k)) \) are obtained by feature extraction using HTK.

Continuous Gaussian mixture-density HMMs are used as acoustic models. A 16-state word-level HMM with no skips over states is used for each of the 11 digits (‘zero’ to ‘nine’ and ‘oh’). Additionally, a three-state silence model with a backward skip from state three to state one is used. Except for the final experiment in Section 5.3, mixtures with 3 Gaussian components are used as output densities for all HMMs. For Baum-Welch training, 20 re-estimation iterations are performed with the HTK tool HERE\textsuperscript{EST}, where a split of mixture components is performed after 10 and 15 iterations to go from 1 to 2 components and to go from 2 to 3 components, respectively. For SPR, a clean-speech HMM trained with 20 Baum-Welch iterations is used as the starting point. Using the stereo data, a reverberant HMM is derived from the clean-speech HMM using the HTK tool HERE\textsuperscript{EST} [13]. The ICEWIND training uses the same clean-speech HMM as starting point. The SFA described by (3) is obtained using the HTK tool HVITÉ. For the estimation of the Gaussian mixture densities, we use our own C++ implementation of the EM algorithm. The mixture components are initialized using k-means clustering. Then, depending on the number of training frames, about 20-40 EM iterations are performed to determine the parameters of the mixture densities.

5.2. Analysis of temporal alignment
First, we verify whether the temporal alignment is improved by ICEWIND compared to Baum-Welch training. Therefore, we determine the Viterbi alignment of the reverberant test data to matched Baum-Welch HMMs and to matched ICEWIND HMMs. These alignments are then compared to the Viterbi alignment of the clean test data to the clean HMMs. The delays of the transitions for the reverberant HMMs relative to the transitions of the clean HMMs are denoted by \( \delta_i \) as illustrated in Figure 1 and are considered as realizations of a random variable \( \Delta \) in the following. Evaluating these delays for the 512 test utterances with an average of about 4 digits per utterance (this means about 2000 HMM transitions), the normalized histograms of Figure 2, showing estimated pdfs \( \hat{p}(\delta) \), are obtained.

The fact that positive delays occur much more frequently than negative delays indicates that the transitions between the reverberant HMMs tend to occur later than the transitions between the clean HMMs. This means that both kinds of matched reverberantly-trained HMMs tend to model the reverberation tail as well. For example, the HMM for digit seven does not only cover those frames where “seven” was uttered by the speaker but also the reverberation tail of seven. Comparing the histograms of the transition delays for the Baum-Welch and the ICEWIND HMMs in Figure 2, it is obvious that the temporal alignment of the ICEWIND HMMs is significantly closer to the clean-speech alignment than that of the Baum-Welch HMMs: The peak around a delay of zero is substantially more pronounced for ICEWIND. Furthermore, the second mode around a delay of 65 frames observed for the Baum-Welch HMMs does not occur with the ICEWIND HMMs. A closer analysis of the data, investigating the delays of the transition from each of the twelve HMMs to each other of the twelve HMMs, explains the second mode for the Baum-Welch HMMs. It is caused by the fact that at the end of each utterance, the reverberation tail, which has a length of 70 frames in our example, is largely modeled by the last digit HMM if Baum-Welch HMMs are used. In contrast, the reverberation tail is correctly modeled by the silence model if ICEWIND is used. These results confirm the idea of ICEWIND that using the clean-speech SFA for training also leads to a better temporal alignment for the test data.

5.3. Recognition results
Table 2 compares the Word Error Rates (WERs) achieved by HMMs trained with different algorithms. It can be seen that the WERs of the HMMs trained on clean speech (a) significantly increase with increasing reverberation. This observation underlines the necessity for adjusting the HMMs to reverberation. The HMMs trained on matched reverberant data (b), (c), (d) perform significantly better in all rooms, regardless of the training method. The performance of SPR is only slightly better than that of Baum-Welch training.

The performance of ICEWIND is marginally better than that of the other training approaches in the moderately reverberant room A, and it is substantially better in the highly reverberant environments of rooms B to E. These results confirm that the more accurate time alignment achieved by ICEWIND also leads to an increased discrimination capability. In other words, if the HMMs model only those frames that correspond to the respective parts of the utterance and if they do not model the
6. Summary and conclusions

A novel approach termed ICEWIND for deriving a matched reverberant HMM from a well-trained clean-speech HMM using stereo data consisting of clean and reverberant feature vectors has been proposed in this paper. While the state-frame alignment is obtained by aligning the clean data to the clean-speech HMM, the parameters of the reverberant HMM’s mixture-densities are estimated by the standard EM algorithm using the reverberant data. Thus, a more accurate temporal alignment is achieved both during training and test so that the discrimination capability of the HMMs is increased. Connected digit recognition experiments confirm the increased recognition performance in comparison to both standard Baum-Welch training and single-pass retraining. Since the effort for training is significantly decreased, ICEWIND should make matched reverberant training attractive for a wider range of applications.

7. Acknowledgments

This work was partly supported by the Deutsche Forschungsgemeinschaft (DFG) under contract number KE 890/4-1.

8. References


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

<table>
<thead>
<tr>
<th>room</th>
<th>$T_{60}$ (ms)</th>
<th>$d$ (m)</th>
<th>$SRR$ (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>300</td>
<td>2.0</td>
<td>~4.0</td>
</tr>
<tr>
<td>B</td>
<td>200</td>
<td>4.1</td>
<td>4.0</td>
</tr>
<tr>
<td>C</td>
<td>900</td>
<td>4.0</td>
<td>0.5</td>
</tr>
<tr>
<td>D</td>
<td>780</td>
<td>2.0</td>
<td>0.4</td>
</tr>
<tr>
<td>E</td>
<td>1300</td>
<td>2.0</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Table 2: Comparison of WERs in % for Baum-Welch and ICEWIND in room C using HMMs with different numbers of mixture components.

<table>
<thead>
<tr>
<th>mixture components</th>
<th>training method</th>
<th>clean A</th>
<th>clean B</th>
<th>clean C</th>
<th>clean D</th>
<th>clean E</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Baum-Welch</td>
<td>0.5</td>
<td>7.4</td>
<td>29.8</td>
<td>47.8</td>
<td>30.5</td>
<td>32.6</td>
</tr>
<tr>
<td>(b) Baum-Welch</td>
<td>-</td>
<td>1.6</td>
<td>5.0</td>
<td>9.2</td>
<td>5.8</td>
<td>7.7</td>
</tr>
<tr>
<td>(c) SPR</td>
<td>-</td>
<td>1.6</td>
<td>5.0</td>
<td>9.0</td>
<td>5.3</td>
<td>6.2</td>
</tr>
<tr>
<td>(d) ICEWIND</td>
<td>-</td>
<td>1.5</td>
<td>3.5</td>
<td>5.4</td>
<td>3.5</td>
<td>3.7</td>
</tr>
<tr>
<td>(e) rel. improv.</td>
<td>-</td>
<td>6.3</td>
<td>30.0</td>
<td>40.0</td>
<td>34.0</td>
<td>43.5</td>
</tr>
</tbody>
</table>

Table 3: Comparison of WERs in % for SPR and ICEWIND in room C with different numbers of mixture components.

<table>
<thead>
<tr>
<th>mixture components</th>
<th>(a) Baum-Welch</th>
<th>(b) Baum-Welch</th>
<th>(c) SPR</th>
<th>(d) ICEWIND</th>
<th>(e) rel. improv.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9.4</td>
<td>8.9</td>
<td>9.0</td>
<td>8.1</td>
<td>7.5</td>
</tr>
<tr>
<td>2</td>
<td>9.1</td>
<td>6.6</td>
<td>5.4</td>
<td>5.3</td>
<td>5.2</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>