Semi-Blind Model Adaptation using Piece-wise Energy Decay Curve for Large Reverberant Environments

Abdul Waheed Mohammed†‡, Marco Matassoni†‡, Harikrishna Maganti‡, Maurizio Omologo‡

†Università degli studi di Trento
‡Center for Information Technology, Fondazione Bruno Kessler Irst
via Sommarive 18, 38123 Trento, Italy
{amwaheed|matasso|maganti|omologo}@fbk.eu

Abstract

This work presents semi-blind acoustic model adaptation based on a piece-wise energy decay curve. The dual slope representation of the piece-wise curve accurately captures the early and late reflection decay that helps in precisely modeling the smearing effect caused due to reverberation. The slopes are estimated in a semi-blind fashion, late reflection slope is estimated blindly by finding the highest likelihood obtained after matching the test features with Gaussian mixture models trained on reverberant data, while the early reflection slope is empirically computed. Adaptation using piece-wise decay curve leads to robust acoustic models consequently improving the recognition performance. The approach is tested on connected digits recognition task in a lecture room with various large reverberation times. The performance is compared with the exponential decay approach and incremental MLLR, where the proposed technique is found to be robust and consistent across all the cases.

Index Terms: reverberation, acoustic model adaptation, GMM

1. Introduction

Current automatic speech recognition (ASR) systems perform very well in close talking scenario but their performance decreases drastically in hands-free environments due to noise and reverberation. Even though several techniques have been proposed to handle additive noise [1], the problem of reverberation remained relatively neglected. Reverberation is a natural acoustic phenomenon caused due to the reflections of the original signal from the walls and objects in the room. To reduce the detrimental effects of reverberation various techniques are used, which can broadly be classified as signal, feature and model based techniques. Although signal and feature based techniques [2] are successful in partially alleviating the effects of convolution distortions, but for large reverberation times (T60) they provide limited gains. Model based techniques such as MAP, MLLR and CMLLR adapt the models by using some data from the target environment [1] and they have shown quite improved performance against noise and reverberation. However, for higher T60s the gains are meager due to the conditional independence assumption of the hidden Markov models (HMM).

In [3, 4], to overcome the conditional independence of HMMs, the reverberation effects in previous frames/states are added to the current frame/state. However, they differ in the reverberation model and the estimation procedure being used to approximate the reverberation effect. In [3], spectral distortion and reflection coefficients of previous frames model the reverberation effects, whereas in [4] exponential energy decay curve (Exp-EDC) representing the energy decay in the room impulse response (RIR) is used. Among these techniques, [4] is closest to the work presented here as it is a blind adaptation method and

we are interested in studying adaptation when no or very little information is available. Despite being robust against small and medium T60, the adaptation in [4] has many limitations. The most important among them are the crude reverberation model and its estimation method. The reverberation model described by Exp-EDC has a monotonous energy decay, where the slope of this decay curve (i.e., T60) is estimated by force matching the adapted HMMs with the test feature vectors. The T60 of the model which provides the highest likelihood is considered as the representative of the target environment. But, due to inaccurate estimation of the reverberation model, the adapted HMMs are not the reliable representative of the target environment. As a result, the estimated T60 values are much smaller than the T60 of the target environment.

To overcome the limitations of Exp-EDC and address the large T60 problem, a better model namely piece-wise energy decay curve (Pw-EDC), which approximates the decay of early and late reflections in a precise manner is proposed. Moreover, to compute T60 which is a essential parameter for this technique, a simple and blind estimation method is also presented. Figure 1 compares the normalized EDC of a lecture room RIR with both Exp-EDC and Pw-EDC for T60 of 830 ms. It is evident from the figure that Pw-EDC models the energy decay more precisely than the Exp-EDC. To obtain accurate estimates of T60, the GMMs trained on reverberant data are used to decode the test feature vectors. The GMM with highest likelihood indicates the estimate of T60.

The organization of the paper is as follows: Section 2 reviews the approach described in [4] and presents the proposed approach. Section 3 provides the details of the experimental setup and the experimental results are discussed in Section 4. Finally, Section 5 presents the summary and future work.
2. EDC based adaptation

In this section, the model adaptation and the estimation of exponential and piece-wise EDC along with the blind estimation of T60 are described.

2.1. Exponential EDC

In a HMM, the acoustic excitation described by the parameters of a single state could be seen in the succeeding state with some attenuation, whereas in a reverberant scenario this attenuation can be observed even in subsequent models. In [4], this decay of acoustic excitation is modeled by the EDC, which is derived as

\[ h(t) = e^{-\frac{\alpha_t}{T_{60}}} \]  

where \( h(t) \) is the RIR.

For the estimation of this curve only T60 is required, which is estimated using maximum likelihood method. In order to compute reverberation contributions from the EDC the average duration of each state is derived as

\[ \text{dur}(S_i) = \frac{1}{1 - P(S_i | S_{i-1}) \cdot t_{end}} \]  

for all states \( S_i \), where \( P(S_i | S_{i-1}) \) is the transition probability to remain in state \( S_i \) and \( t_{end} \) is the frame shifting time.

The reverberation contribution \( \alpha_i \) of each state is estimated by integrating the squared RIR over the time segment of each state as

\[ \alpha_i = \int_{t_i}^{t_{i+1}} h^2(t) \, dt \]  

where \( t_i \) and \( t_{i+1} \) are start and end times of each state respectively.

The adaptation of the energy parameter and MFCCs of the current state can be performed by adding the reverberation contributions of the current state and the preceding states. Since the adaptation is defined in Mel-spectral domain, the cepstral coefficients are transformed back to the Mel-spectral domain and then the adaptation is performed as

\[ \hat{E}(S_i) = \sum_{j=1}^{i} \alpha_{i,j} \cdot E(S_j) \]  

\[ |\hat{X}_k(S_i)|^2 = \sum_{j=1}^{i} \alpha_{i,j} \cdot |X_k(S_j)|^2 \]  

where \( k \) is the filter bank index, \( X_k \) is the clean power density spectra, \( E \) is the clean energy parameter, \( \hat{X}_k \) is the adapted spectra, \( \hat{E} \) is the adapted energy parameter. After the adaptation, the spectral parameters are again transformed back to MFCCs. Further details are described in [4].

2.2. Piece-wise EDC

In order to model Pw-EDC, the boundaries of early and late reflections needs to be defined. However, in the literature, there are no definite boundaries defined for them. According to [6], the reflections between 50 ms after the arrival of direct sound and when the sound pressure level drops below 40dB have the most detrimental effect on ASR accuracy. Therefore, in the proposed approach 50ms after the arrival of direct sound is considered as the early reflection time and the reflections arriving after this time are considered as late reflections.

The initial part of the Pw-EDC is modeled by the combination of linear and power functions which are derived as

\[ f(t) = mt + c \]  

where \( m \) is the slope and \( c \) is the y-intercept, and

\[ f(t) = t^a \]  

where \( a \) is the power exponent.

The parameters of these functions are empirically computed. Since the later part in EDC is fairly approximated by equation (1), it has been retained in the proposed approach for modeling the late reflection decay. Hence, the only parameter needed to create the late reflection decay is T60:

\[ h(t) = e^{-\frac{60(a10)}{T_{60}}} \]  

The piece wise curve can then be derived as

\[ F(t) = \begin{cases} f(t) \cdot f_p(t) & 0 \leq t \leq 50ms \\ f_p(t) & t > 50ms \end{cases} \]  

2.3. Blind estimation of T60

The proposed approach is blind since it does not require any prior knowledge like speech/silence segmentation or any close-talking reference signal. The estimation procedure is as shown in figure 2. Initially, the reverberant signals are obtained by convolving the clean signals taken from the training set of AU-RORAS [7] corpus with the RIR. In order to have a fair evaluation, the RIR from the office room obtained through the SIREAC [8] tool is used. Using SIREAC tool, T60 of the RIR of the office room is varied from 200, 300, …, 900. Then the clean signals are convolved with these modified RIRs to obtain 8 sets of reverberant corpora respectively.

Features from the reverberant signals are extracted as described in section 3. GMMs with 32 mixtures representing each digit are trained on reverberant features using the standard Expectation and Maximization algorithm to obtain 8 sets of GMM \( T_{60}(k) \), where T60 represents the T60 of the corpus used and \( k \) corresponds to the digits in the corpus. Viterbi decoding of the test features is performed using these GMMs. The likelihoods of \( GMM_{T_{60}(k)} \) are compared and the highest among them provide the closest match to the environment, thus the T60 of that \( GMM_{T_{60}(k)} \) is taken as the estimated T60 for the utterance. The results of the estimation procedure along with the analysis are provided in section 4.
3. Experimental setup

To evaluate the proposed method, the RIRs taken from the AIR database are used [9]. It contains binaural RIRs with sampling rate of 48 kHz measured in four rooms: studio booth, office room, meeting room and lecture room. The measurements are carried out with and without dummy head. Moreover, in each room, RIRs are captured at different distances having different T60s, providing a wide range of scenarios where hands-free devices can be used. The specifications of the RIRs are shown in Table 1. For each case i.e., T60, the database contains the RIRs for left and right channel, with and without dummy head, thus a total of 24 RIRs are available for the lecture room.

To generate the reverberant corpora, clean signals are taken from the clean set of AURORA5, upsampled to 48 kHz and then convolved with the lecture room RIRs and finally down sampled to 8 kHz to obtain 24 reverberant corpora. The features are calculated by pre-emphasizing the signal with a factor of 0.95. The short segments of speech are extracted using a hamming window of 25ms with a frame shift of 10ms. Spectral analysis on frames is performed with 256 point DFT (Discrete Fourier Transform). Mel-spectrum is calculated by applying a Mel-filter bank having 24 band-pass filters in the range from 200 Hz to 4000 Hz on the DFT spectrum. MFCCs are obtained from the log mel-spectrum by applying DCT (Discrete Cosine Transform). In the experiments, C0 is needed only for transforming the cepstral coefficients to the spectral domain during adaptation. For the recognition, static coefficients with energy parameter augmented with delta coefficients are used. The word models consist of 16 emitting states and 4 Gaussian mixtures per state which represents the digits, whereas silence model has 3 states with 4 Gaussian mixtures per state. Training and testing of the models are performed on the HTK toolkit [10] by using the whole corpus. For the experiments, the means of the static coefficients are adapted. The clean models are adapted with the estimated T60 value using exponential and piece-wise adaptation method. The adapted models are then used for recognition.

Figure 3 describes the piece-wise adaptation method. Initially, T60 is estimated using the GMM for each utterance as described in section 2.2, then the mean of the T60 values for the whole test corpus is used to create the Pw-EDC as described in section 2.2. The duration of each state is calculated from the transition probabilities of the clean models. Reverberation contributions are estimated by integrating the Pw-EDC for each state duration. Mel-spectrum of the clean models are adapted by adding the reverberation contributions of previous states as described in [4]. After the adaptation, the mel-spectrum coefficients are transformed back to the cepstral domain, thus obtaining the Pw-EDC adapted models.

4. Experimental results

In order to test the efficacy of our approach, extensive experiments are performed using the connected digit setup. Initially, the results of the estimation of T60 are discussed, then the performance of exponential and piece-wise adaptation with unsupervised maximum likelihood linear regression (IMLLR) are compared. Finally, the influence of T60 over the adaptation techniques is evaluated.

4.1. Comparison of T60 estimation

In this section, the comparison of T60 estimation using trained GMMs and exponentially adapted HMMs is discussed. Figure 4 shows the results of the experiment. Initially, the trained GMMs and exponentially adapted HMMs is discussed. Figure 4 shows the results of the experiment. Initially, the trained GMMS are used to decode the lecture room test features for each utterance, the model which provide the highest likelihood is taken as the closest match of the environment and its T60 value is considered the estimated T60 value. The estimated values for all the utterances in each test set are accumulated and the average of these values estimated over all the 24 test sets is shown in figure 4. The mean T60 value estimated over all the sets and over all the utterances is 754 ms, which is very close to the overall mean T60 i.e., 780 ms estimated over the lecture room RIRs using Schroeders method [5].

The same procedure is followed to estimate T60 using exponentially adapted HMMs. The average of the estimated T60 values is plotted in the figure 4(b). It is evident from the figure that most of the estimated values are considerably lower than the actual T60 values. Thus, it clearly demonstrates that the adapted HMMs are not properly adapted due to the crude reverberation model, whereas the trained GMMs are more effective in estimating T60 even without any processing of speech.

4.2. Comparison with standard adaptation technique

The aim here is to study blind adaptation and compare it against a standard adaptation technique i.e., IMLLR. Initially, the clean models are adapted using the IMLLR, exponential and piece-wise adaptation techniques. IMLLR adaptation is performed by adding the reverberation contributions of previous states as described in [4]. After the adaptation, the mel-spectrum coefficients are transformed back to the cepstral domain, thus obtaining the Pw-EDC adapted models.

Table 1: Lecture room impulse responses specifications

<table>
<thead>
<tr>
<th>Distance (m)</th>
<th>2.25</th>
<th>4</th>
<th>5.56</th>
<th>7.1</th>
<th>8.68</th>
<th>10.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>T60 (ms)</td>
<td>700</td>
<td>720</td>
<td>790</td>
<td>800</td>
<td>810</td>
<td>830</td>
</tr>
</tbody>
</table>
after each utterance using a transform with a regression tree having 4 nodes. For exponential and piecewise adaptation estimated T60 values are used. Other parameters for piecewise adaptation are empirically determined. For each case i.e., T60, there are 4 different sub-cases e.g., the results of right and left channel with and without dummy head. In order to provide succinct results, the average of these 4 sub-cases for each case is computed.

The results are shown in Figure 5. The exponential adaptation technique provides some marginal improvements despite incorrectly estimating the T60 values (for analysis see section 4.3). Although, in most cases IMLLR is able to outperform the exponential adaptation, the improvements are not consistent throughout all the cases. However, the piece-wise adaptation provides substantial decrease in the WER across all the cases and in almost all the cases improving upon the IMLLR which clearly demonstrates the robustness and consistency of our approach. It is apparent from the results that the accurate estimation of T60 helped in computing a precise Pw-EDC which was effective in modeling the reverberation accurately, thus improving the recognition performance.

4.3. Influence of reverberation time

In this section, robustness of the exponential and piece-wise adaptation against incorrect T60 value is evaluated. Initially, the T60 values are sampled as 50, 100, . . . , 1000 ms and then clean models are adapted using exponential and piece-wise adaptation. The resulting adapted models are tested on the reverberant corpora of a lecture room. Figure 6 shows the results of the experiment for two cases of the lecture room, where the T60 is 700 and 830 ms respectively. It is evident that the exponential adaptation is sensitive to the choice of T60. For lower T60 it gives the best results, whereas for higher T60, due to the overestimation of reverberant contributions the adaptation is performed incorrectly, consequently the WERs are always higher than the unmatched results. A similar trend is observed across all the cases. Moreover, the best results in the case of exponential adaptation is obtained at a value other than the actual T60 of the channel.

In the case of piece-wise adaptation, due to precise modeling of energy decay the technique shows decrease in the WER, providing robustness against incorrect estimation of T60. However, when there is a large difference between the T60 of the channel and the T60 used for adaptation it shows a small increase in WER indicating the incorrect T60 values. The most interesting aspect of this technique is it provides the best results at the same T60 value as of the channel. Similar trend has been observed for all the cases.

5. Summary and future work

This paper has presented a semi-blind acoustic model adaptation technique for large reverberant environments. The adaptation is performed using a Pw-EDC which accurately models the early and late reflections decay in a channel. Connected digits recognition experiments are performed for various T60s in a lecture room of the AIR database. The results are compared with the exponential adaptation and IMLLR. The results obtained are significantly better than the exponential model and IMLLR. In the future, the approach will be extended to become totally blind by estimating all the parameters necessary to create the Pw-EDC. Moreover, to increase the robustness and applicability to real scenarios, the approach will be extended by incorporating noise modeling and adapting the dynamic coefficients (i.e., first and second order coefficients).

6. Acknowledgments

We express gratitude to Prof. Hirsch, Niederrhein University, Germany for providing us help with the scripts.

7. References