ACOUSTIC MODEL ADAPTATION USING PIECE-WISE ENERGY DECAY CURVE FOR REVERBERANT ENVIRONMENTS

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

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

ABSTRACT

This work presents 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 which helps in precisely modeling the smearing effect caused due to reverberation. 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 different rooms with various reverberation times. The performance is compared with the exponential decay approach and incremental MLLR, where the proposed approach was more effective. Moreover, the combination of piece-wise adaptation with incremental MLLR is also studied and the combination is instrumental in improving the performance with respect to incremental MLLR.

Index Terms— reverberation, robust speech recognition, acoustic model adaptation

1. INTRODUCTION

Current automatic speech recognition (ASR) systems perform very well in close talking scenario but their performance decreases drastically in hands-free scenario due to noise and reverberation. Even though many effective techniques have been proposed to tackle 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.

Signal based techniques [2] mostly focus on improving the perceptual quality of the signal by maximizing the signal to noise ratio (SNR) whereas feature based techniques [3] extract robust features from the speech which are immune to the effects of the environment. Nevertheless, these techniques are successful in partially alleviating the effects of convolution distortions but for large reverberation times they provide limited gains. Model based techniques like Maximum a-posteriori (MAP) [4], Maximum likelihood linear regression (MLLR) [5] and Constrained maximum likelihood linear regression (CMLLR) [6] aim to reduce the mismatch between training and testing conditions by adapting the models using some data from the target environment. Although, they have shown quite improved performance against the noise and reverberation, for higher reverberation times the gains are meager due to the conditional independence assumption of the hidden Markov models (HMM).

Recently [7, 8, 9] have attempted to overcome the conditional independence of HMMs. The basic idea in [7, 8] is to estimate the reverberation effects in the preceding frames and add it to the current frame, whereas in [9], the reverberation effects are estimated from the previous states and added to the current state. Moreover, these techniques also differ in the reverberation model and the estimation procedure being used to approximate the reverberation effect. In [7], the spectral distortion and the reflection coefficients of the previous frames model the reverberation effects, whereas in [8], only the reflection coefficients of the previous frames are used; while in [9], the exponential energy decay curve (Exp-EDC) which represents the energy decay in the room impulse response (RIR) is used. Among these techniques, [9] is closest to the work presented in this paper as it is a blind adaptation procedure and we are interested in studying the adaptation when no information or very little information is available.

The adaptation in [9] is based on the Exp-EDC which assumes a monotonous energy decay in the channel. However, in [10] it is shown that the EDC has dual slope representing the early and late reflections decays. Therefore, in this work we use a better model for creating the piece-wise energy decay curve (Pw-EDC) which approximates the decay of both reflections in a precise way. Clean models are adapted by using the Pw-EDC and tested under different rooms with several different reverberation times (T60). Moreover, the combination of piece-wise adaptation with unsupervised MLLR commonly referred to as incremental MLLR (IMLLR) is also studied.

The organization of the paper is as follows: Section 2, re-

views the approach presented in [9]. Section 3 presents the proposed approach. Section 4 provides the details of the experimental setup and the experimental results are discussed in Section 5. In Section 6, the summary and the future work are presented.

2. ADAPTATION USING EXPONENTIAL ENERGY DECAY CURVE

In a HMM, the acoustic excitation described by the parameters of a single state could be seen in the succeeding states with some attenuation. In a reverberant scenario, this attenuation can be observed even in the succeeding models. In [9], this decay of acoustic excitation is modeled by the EDC derived as

$$h^2(t) \sim e^{-\frac{6ln(10)}{T_{60}} \cdot t} \tag{1}$$

where $h(t)$ is the RIR.

In order to estimate this curve, only T60 is needed which is estimated using a maximum likelihood method. To estimate the reverberation contribution in each state, first the average duration of each state is derived as

$$dur(S_i) = \frac{1}{1 - P(S_i \mid S_i)} \cdot t_{shift} \tag{2}$$

for all states S_i, where $P(S_i \mid S_i)$ is the transition probability to remain in state S_i and t_{shift} is the frame shifting time. Then the reverberation contribution α_i of each state is estimated by integrating the squared RIR over the time segment of each state as

$$\alpha_{i,1} = \int_{t_s(S_i)}^{t_e(S_i)} h^2(t)dt \tag{3}$$

where t_s and t_e are starting and ending times of each state respectively.

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

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

$$|\hat{X}_k(S_i)|^2 = \sum_{j=1}^{i} \alpha_{i,j} \cdot |X_k(S_j)|^2 \tag{4}$$

where k is the filter bank index and X_k is the clean power density spectra and E is the clean energy parameter; similarly $\hat{X}_k$ and $\hat{E}$ are the adapted spectra and the adapted energy parameter respectively. After the adaptation, the spectral parameters are again transformed to MFCCs. For more details refer to [9].

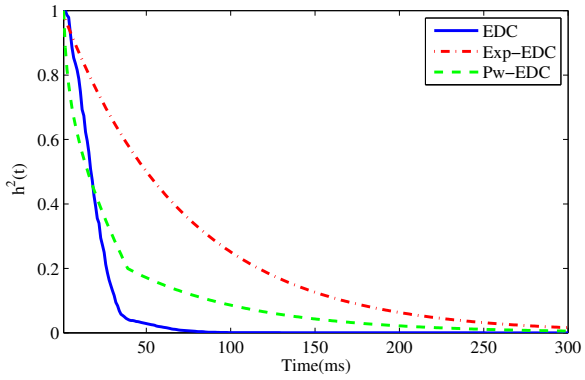

Fig. 1. Comparing energy decay curve of the office room RIR having T60 $\sim$ 500ms with exponential and piece-wise energy decay curves of T60=500ms

2.1. Limitations of the exponential adaptation

In the exponential adaptation approach, the Exp-EDC describes the energy decay in the channel using a single and monotonous decay. However, according to [10] the reverberant energy decays in two phases; in the initial phase it decays sharply due to high energy sparse reflections while in the later phase the energy decays smoothly due to the low energy dense reflections. This effect is illustrated in figure 1 which shows the normalized EDC of the RIR captured in an office room and Exp-EDC estimated using equation (1) both having T60 of 500 ms. It is evident, the Exp-EDC is not able to approximate the initial and some later part of the RIR's EDC consequently leading to overestimation of the reverberation contribution in the states. Hence, to capture accurate representation of the reverberation it is necessary to model the EDC with dual slope.

In the exponential adaptation approach, T60 plays a fundamental role because the Exp-EDC is created using it. However, it is estimated by force matching the adapted models with the feature vectors using a maximum likelihood criteria and it has been reported in [9] that the estimated T60 values vary a lot. Therefore, in a preliminary investigation we have also estimated the T60 using the same procedure and found that most of the estimated values are much smaller than the actual T60 of the environment. Hence, to remove this anomaly in our experiments we have used the actual T60 of the environment instead of depending upon the inaccurately estimated values.

3. ADAPTATION USING PIECE-WISE ENERGY DECAY CURVE

In order to model the Pw-EDC, the boundaries of early and late reflections need to be defined. However, in the literature, there are no definite boundaries defined for them. According to [11], the reflections between 50 ms after the arrival of the

direct sound and when the sound pressure level drops below 40dB have the most detrimental effect on ASR accuracy. Moreover, [12] has reported that the late reflections which occurs between 100 ms and 300 ms after the direct signal are the most harmful for classification accuracy. Therefore, in our experiments we have chosen 50 ms after the arrival of the direct sound as the early reflection time and the reflections arriving after this are considered as the late reflections. Using these boundaries we have modeled the Pw-EDC curve as follows: The initial part of the Pw-EDC is modeled by the combination of linear and power function derived as

$$f_l(t) = m \cdot t + c \tag{5}$$

where m is the slope and c is y the intercept, and

$$f_p(t) = t^a \tag{6}$$

where a is the power exponent. The parameters of these functions are empirically computed. As the later part in EDC was fairly approximated by equation (1), we have retained the same modeling in our approach for the late reflections decay. The only parameter needed to create the late reflection decay curve is T60 which is already known.

$$f_e(t) \sim e^{-\frac{6ln(10)}{T_{60}} \cdot t} \tag{7}$$

The piece wise curve can then be derived as

$$\begin{aligned} F(t) &= f_l(t) \cdot f_p(t) \quad 0 \leq t \leq T \\ &= f_e(t) \qquad\quad t > T \end{aligned} \tag{8}$$

where T is 50 ms . Figure 1 compares various energy decay curves where the Pw-EDC having early reflection boundary as 50 ms is displayed. It is apparent from the figure 1, the Pw-EDC models the early and late reflections more accurately than the simple Exp-EDC.

4. EXPERIMENTAL SETUP

To evaluate the proposed method, the reverberant data is generated for office and living rooms with the reverberation times ranging from 200 to 900 ms. The RIR for each room is obtained from the web interface of SIREAC tool [13]. This tool provides several options to modify the RIR in terms of selecting rooms and reverberation time. First, the clean corpus is obtained after down sampling the TIDIGITS corpus [14] to 8 kHz, then the reverberant corpora is created by convolving the RIRs with clean signals. In this manner, we have created 8 sets of reverberant corpora corresponding to the T60 of 200, 300 , . . . , 900 ms for each room.

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 25 ms with a frame shift of 10 ms. Spectral analysis on frames is performed with 256 point

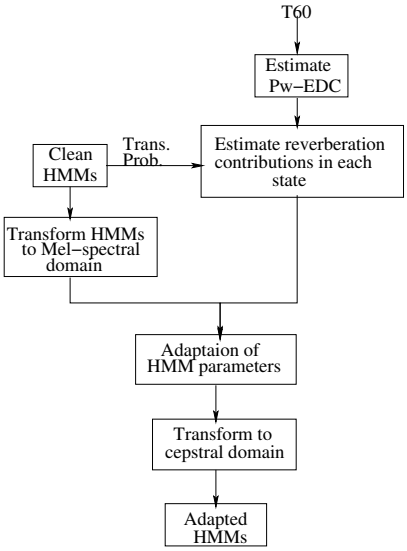

Fig. 2. Piece-wise adaptation of acoustic models

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 our experiments, zeroth cepstral coefficient (C0) is needed only for transforming the cepstral coefficients to the spectral domain during adaptation. For recognition, static coefficients with energy parameter augmented with delta coefficients are used.

The word models consists 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 HTK toolkit [15] by using the whole corpus. In this contribution, only the static coefficients of the means are adapted. The clean models are adapted with the fixed T60 value using the exponential and piece-wise adaptation method. The adapted models are then used for recognition.

Figure 2 presents the piece-wise adaptation method. Initially, the Pw-EDC is estimated as described in section 3. Then the duration of each state is calculated from the transition probabilities of clean models. Reverberation contributions are estimated by integrating the Pw-EDC for each state duration. Adaptation of the clean models is performed by adding the reverberation contributions of the previous states to the current state as described in [9]. After the adaptation, the models are transformed back to the cepstral domain thus obtaining the Pw-EDC adapted models.

5. EXPERIMENTAL RESULTS

In order to test the efficacy of our approach extensive experiments are performed using connected digit setup. Initially, the influence of T60 is assessed and then we compare the perfor-

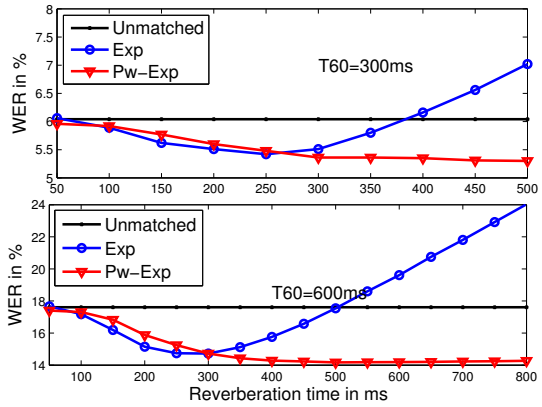

Fig. 3. Influence of T60 over the exponetial and piece-wise adaptation techniques

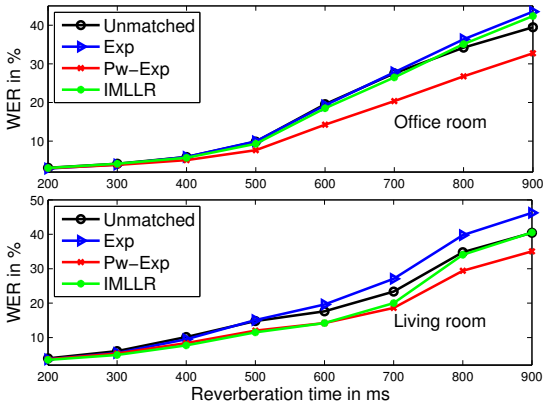

Fig. 4. Comparison of exponential and piece-wise adaptation with IMLLR

mance of exponential and piece-wise adaptation with IMLLR. Finally, the combination of piecewise adaptation and IMLLR is studied.

5.1. Influence of reverberation time

Generally, T60 is estimated from the RIR but many techniques have also been proposed to estimate it blindly from the signals. However, the blind estimation techniques does not provide the exact T60 values . In this scenario, we have investigated the robustness of the exponential and piece-wise adaptation when an incorrect T60 value is provided to the algorithm.

Initially, the T60 values are sampled in the steps of 50 ms starting from 50 ms till 1000 ms. The clean models are adapted using these T60 values by exponential and piece-wise adaptation. The resulting models are then tested on the reverberant corpora. Figure 3 shows the results of this experiment for only two cases of living room where the T60 is 300 and 600 ms respectively. From the results, it is evident that exponential adaptation is very sensitive to the choice of T60. For lower T60 it gives the best results whereas for higher T60 due to the over-estimation of reverberant contributions the adaptation is performed incorrectly, hence the word error rates (WER) are always higher than the unmatched results. Similar trend is observed for all the T60 values across both the rooms. Moreover, the best results in the case of exponential adaptation is obtained at a value other than the original T60 of the channel.

In the case of piece-wise adaptation, due to the precise modeling of energy decay the technique shows decrease in the WER, thus 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 reverberation times across both the rooms.

5.2. Comparison with standard adaptation technique

The aim here is to study the adaptation when no information or some information about the environment is available like T60. Therefore, we have chosen to compare our technique with IMLLR . Initially, clean models are adapted using IMLLR, exponential and piecewise adaptation techniques. The IMLLR adapation is performed after each utterance using a global transform. For exponential and piecewise adaptation T60 values are provided. Other parameters for piecewise adaptation are empirically determined. The results of this experiment for both rooms are shown in figure 4.

In the office room scenario, at lower T60s the relative gains due to adaptation is meager for all the techniques. Moreover, as the reverberation time increases the exponential adaptation is showing worse performance than the unmatched case. Nonetheless, the robustness of piece-wise adaptation is clearly visible particularly for higher T60 values where even IMLLR is showing dismal performance.

In the living room scenario, the trend at lower T60s is similar to the office room but at higher reverberation times piece-wise adaptation outperforms other techniques. Finally, the piece-wise adaptation due to the accurate modeling of the reverberation effects has shown to be robust and consistent across both the rooms particularly at higher T60s.

5.3. Combination of adaptation techniques

In order to obtain robust models, multiple adaptations techniques could be used for example the combination of MLLR and MAP adaptation is well known. Therefore, we have attempted to study the combination of IMLLR with exponential and piece-wise adaptation. In the previous section, it has been observed that IMLLR is not particulary effective in large T60

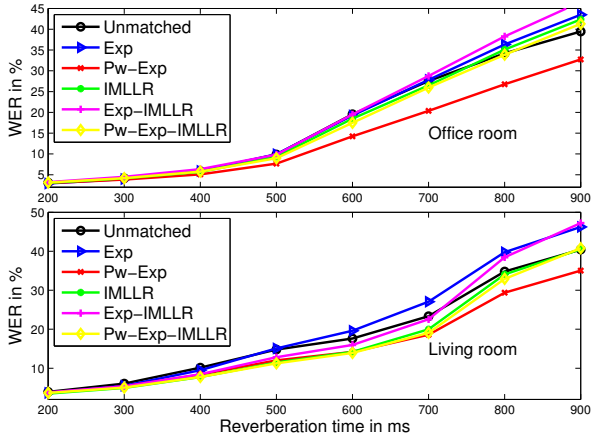

Fig. 5. Comparison of Exp-IMLLR and Pw-Exp-IMLLR adaptation with IMLLR

scenario. Therefore, our aim is to improve the IMLLR performance by first adapting the models using exponential or piecewise adaptation and then performing IMLLR over these models.

Figure 5 presents the results of this study. The combination of exponential and IMLLR adaptation (Exp-IMLLR) is not beneficial except at lower reverberation times because the exponential adaptation has already adapted the models in an incorrect way. As a result, the prior adaptation is not helpful in creating accurate transforms. However, in the low reverberation scenario when the exponential adaptation is showing some improved performance, the IMLLR is benefited from the exponential adaptation and the WERs are reduced. For the piece-wise and IMLLR adaptation (Pw-Exp-IMLLR), IMLLR is benefited due to the piece-wise adaptation in all the cases across both the rooms.

6. SUMMARY AND FUTURE WORK

This paper has presented an improved acoustic model adaptation technique for reverberant environments. The adaptation is performed by using a Pw-EDC which accurately models the early and late reflections decay in a channel. Connected digits recognition experiments have been performed in different rooms for various reverberation times and the results are compared with the approach in [9] and IMLLR. The results obtained are significantly better than the exponential model and IMLLR. Finally, the combination of IMLLR with exponential and piece-wise adaptation is also studied and it is found that piece-wise adaptation helps in improving IMLLR performance. In the future, we would like to address the issue of estimation of the parameters for creating the Pw-EDC. A possible approach could be to use some amount of reverberant data. Moreover, to increase the robustness and applicability to real scenarios we would like to extend the approach by incorporating noise modeling and adapting the dynamic coefficients (i.e first and second order coefficients).

7. ACKNOWLEDGMENTS

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

8. REFERENCES

[1] J. Droppo and A. Acero, "Environmental robustness," *Springer Handbook of Speech Processing*, vol. II, pp. 653–679, 2008.

[2] P.A. Naylor and N.D. Gaubitch, "Speech dereverberation," *Springer*, 2010.

[3] B. Atal, "Effectiveness of linear prediction characterisitics of the speech wave for automatic speaker identification and verification," *Journal of the Acoustic Society of America*, vol. 55(6), pp. 1304–1312, June 1974.

[4] J.L. Gauvin and C.H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observation of Markov chains," *IEEE Trans. for Speech and Audio Process.*, vol. 2(2), pp. 291–298, April 1994.

[5] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9(2), pp. 171–185, April 1995.

[6] V. Digalakis, D. Rtischev, L. Neumeyer, and S. Edics, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Trans. on Speech and Audio Process.*, vol. 3(4), pp. 357–366, 1995.

[7] T. Takiguchi and M. Nishimura, "Acoustic model adaptation using first order linear prediction for reverberant speech," *ICASSP*, vol. 1, pp. 869–872, 2004.

[8] C.K. Raut, T. Nishimoto, and S. Sagayama, "Model adaptation by state splitting of HMM for long reverberation," *Proc. of Interspeech*, pp. 277–280, 2005.

[9] H.-G. Hirsch and H. Finster, "A new approach for the adaptation of HMMs to reverberation and background noise," *Speech Communication*, vol. 50(3), pp. 244–263, September 2008.

[10] M.R. Schroeder, "New method of measuring reverberation time.," *Journal of the Acoustic Society of America*, vol. 37(3), pp. 409–412, March 1965.

[11] I. Tashev and D. Allred, "Reverberation reduction for improved speech recognition," *Proc. of HSCMA*, 2005.

[12] R. Petrick, K. Lohde, M. Wolff, and R. Hoffmann, "The harming part of room acoustics in automatic speech recognition," *Proc. of Interspeech*, pp. 1094–1097, 2007.

[13] H.-G. Hirsch and H. Finster, "The simulation of realistic acoustic input scenarios for speech recognition systems," *Proc. of Interspeech*, pp. 2697–2700, 2005.

[14] R.G. Leonard, "A database for speaker-independent digit recognition," *Proc. of ICASSP*, vol. 3, pp. 42.11, 1984.

[15] S. Young, et al., "The HTK book (version 3.4.1), Cambridge University Engineering Department, http://htk.eng.cam.ac.uk.," 2009.