EUSIPCO 2013 1569746465

ADAPTATION OF HMM DYNAMIC PARAMETERS IN REVERBERANT ENVIRONMENT

Jinkyu Lee, Hyunson Seo, and Hong-Goo Kang

Department of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea

ABSTRACT

This paper presents a new adaptation method for HMM-based automatic speech recognition system in a reverberant environment. Unlike the conventional approach that estimates dynamic mean vectors by adopting a spline interpolation technique, the proposed algorithm uses the transform derived by the mathematical property. Additionally, we introduce the adaptation for covariance matrices with the domain conversion process induced by log-normal distribution, because the statistical parameters are affected by not only mean vectors but also covariance matrices. Consequently, all statistical parameters in HMM can be adapted by the exact same transform structure. Experimental results show that the proposed method improves the recognition rate, in spite of having much simple adaptation process. Also it is robust to the estimation error that is unavoidable while extracting the reverberation time related parameters.

Index Terms— Robust automatic speech recognition, dereverberation, model adaptation.

1. INTRODUCTION

Although the performance of automatic speech recognition (ASR) system has been increased significantly, the application area of the system is still limited because the performance drops if the system is used in reverberant or noisy environments. The reason can be found from the mismatch caused by the variation of acoustical environments between input speech features and trained acoustic models. In order to reduce the mismatch, a variety of approaches have been proposed. Spectral subtraction [1] or inverse filtering [2] are typical examples to improve the robustness in reverberant environments. Those methods suppress the late reverberation component effectively, whereas they are still inappropriate for the ASR system because of inevitable spectral distortion.

For such reason, model adaptation approaches for additive and convolutive noise environment have been proposed. The approaches estimate the corrupted hidden Markov model (HMM) [3] from the clean acoustic model by using parallel model combination (PMC) [4] and vector Taylor series (VTS) [5]. Although these methods improve the recognition performance somehow, the adaptation algorithm cannot be applied directly to reverberant environments because the length of room impulse response (RIR) is usually much longer than that of analysis window.

Recently, a model adaptation method for reverberant environment has been introduced. This method is to adapt static spectral parameters using an approximated exponential function of RIR [6]. Moreover, it has been extended to adapting dynamic parameters by introducing a spline interpolation technique [7]. Although the adaptation method improves word error rate (WER) somehow, it requires high computational complexity. Besides, the adaptation method including both static and dynamic parameters even results in lower performance than the one including only static parameters if the reverberation time parameter (i.e. T_{60}) is improperly estimated.

This paper proposes an efficient adaptation method for dynamic parameters of HMM. The proposed algorithm derived by the property of *differentiation of convolution* enables that the dynamic parameters can be also included into the same form of adaptation formula to the one for static parameters. In other words, both static and dynamic parameters can be processed simultaneously. Moreover, utilizing the fact that the probability density function is still Gaussian after performing discrete cosine transform (DCT), the property of log-normal distribution can be applied while transforming statistical parameters of HMM. To evaluate the effectiveness of the proposed algorithm, connected digits recognition experiments are performed in six different room environments. The proposed method shows higher word accuracy than conventional methods in all the tested reverberant environments.

The rest of paper is organized as follows. In Section 2, an overview of reverberation effect in the parametric domain of HMM is provided, then conventional adaptation algorithms are briefly reviewed. In Section 3, the proposed adaptation algorithm is described in detail. The performance evaluation results are given in Section 4, and conclusion follows in Section 5.

2. HMM MODEL ADAPTATION FOR REVERBERANT ENVIRONMENT

2.1. Long convolutional distortion

Channel distortion is typically modeled by a convolution between speech signal $x[n]$ and channel characteristic $h[n]$ in

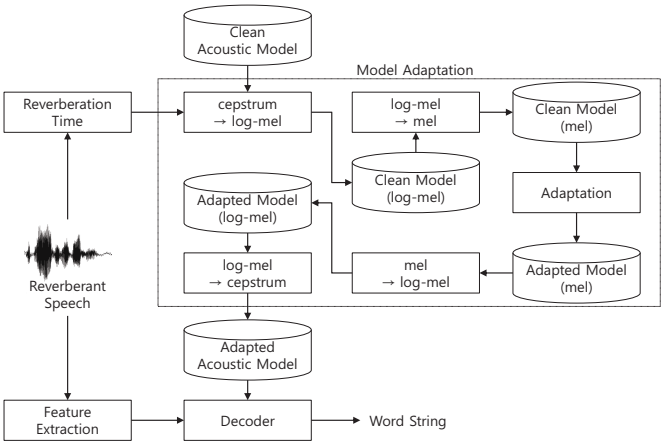

Fig. 1. Schematic diagram of recognition system for reverberant environment

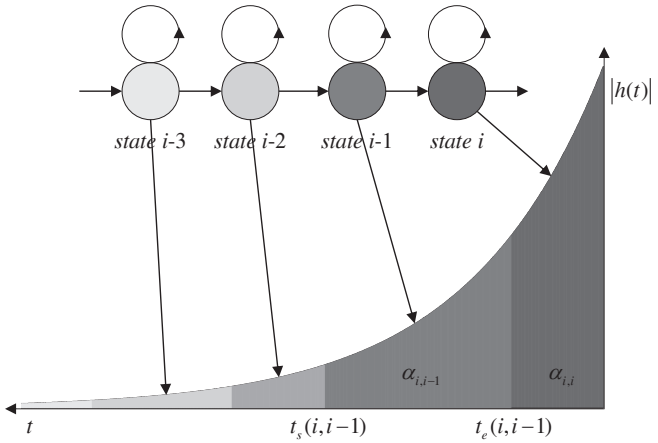

Fig. 2. Estimation of weighting coefficients $\alpha_{i,i-j}$

time domain. In frequency domain, it can be interpreted by a multiplication of two terms given as follows:

$$Y(l,k) = H(l,k)X(l,k), \tag{1}$$

where l and k are the frame and frequency bin index respectively. However, when the length of window for short time Fourier transform (STFT) is much shorter than that of RIR, the convolution in time domain cannot be represented as a multiplication in frequency domain. Instead, the reverberant signal is usually approximated by a convolution in frequency domain because it can be represented by a superposition of delayed frame sequences. Therefore, the approximated reverberant signal in STFT domain is given by

$$Y(l,k) \approx \sum_{i=0}^{L_F-1} H(i,k)X(l-i,k), \tag{2}$$

where L_F is the frame length of RIR [8].

2.2. Adaptation of static parameters

Since the adaptation process is performed in mel spectral domain, the acoustic model needs to be converted from cepstral domain to mel spectral domain. The detailed conversion process is depicted in Fig. 1. The concept of the adaptation is that reverberant spectral parameters can be approximated by weighted sum of the spectral parameters in the current and previous states. To simplify the adaptation process, probability density functions in all states except current state are considered as a single Gaussian mixture. Therefore, the averaged mean vector can be obtained by

$$\bar{\boldsymbol{\mu}}_x^{\mathrm{mel}}(i-j) = \sum_{m=1}^{M} w_m \boldsymbol{\mu}_x^{\mathrm{mel}}(i-j,m), \tag{3}$$

where w_m is the weight value for mth mixture component. Fig. 2 represents the weighting coefficients which indicate the amount of attenuation derived by RIR. If the RIR is modeled as an exponentially decaying function, the coefficient of $(i-j)$th state based on the ith state is calculated as follows:

$$\alpha_{i,i-j} = \frac{\int_{t_s(i,i-j)}^{t_e(i,i-j)} |h(t)|\,dt}{\int_0^{\infty} |h(t)|\,dt}, \tag{4}$$

where $t_s(i,i-j)$ and $t_e(i,i-j)$ are the start and end time of $(i-j)$th state calculated from the ith state, respectively [8]. The start and end time of each state are usually estimated by self transition probabilities in HMM (See [7] for details). Then, the spectral mean vector of reverberant speech can be approximated by

$$\boldsymbol{\mu}_y^{\mathrm{mel}}(i,m) = \alpha_{i,0}\boldsymbol{\mu}_x^{\mathrm{mel}}(i,m) + \sum_{j=1}^{i} \alpha_{i,i-j}\bar{\boldsymbol{\mu}}_x^{\mathrm{mel}}(i-j). \tag{5}$$

Finally, the adapted acoustic model in cepstral domain is obtained by taking a logarithmic function and multiplying a DCT kernel.

2.3. Adaptation of dynamic parameters

It is known that the dynamic characteristic of acoustic model also changes in the reverberant environment. In the previous research [7], dynamic parameters were estimated from the continuous spectral trajectory by the weighted sum between interpolated clean model and the average differences given as follows:

$$\boldsymbol{\mu}_{\Delta y}^{\log}(i) = \boldsymbol{\mu}_{\Delta x}^{\log}(i) + \beta \cdot \bar{\boldsymbol{\mu}}_{\mathrm{diff}}^{\log}\left(\frac{t_s(i,0)+t_e(i,0)}{2}\right), \tag{6}$$

where $\bar{\boldsymbol{\mu}}_{\mathrm{diff}}^{\log}$ is defined as the difference between adapted log-mel spectrum $\bar{\boldsymbol{\mu}}_{\Delta y}^{\log}$ and clean log-mel spectrum $\bar{\boldsymbol{\mu}}_{\Delta x}^{\log}$.

3. PROPOSED ALGORITHM

3.1. Dynamic parameters adaptation using the time differentiation of convolution

In the conventional algorithm [7], plenty of computational power is used for adapting dynamic parameters. It is also found that the performance degrades if the reverberation time parameter is not properly estimated. We thought that the statistical characteristic of dynamic parameters can be derived by another form similar to the static parameters because dynamic parameters are calculated by the weighted sum of neighboring static parameters.

The proposed adaptation method for dynamic parameters is derived by the time *differentiation of convolution*, which has the following property for all continuous time signals f and g, where t indicates the time index.

$$\begin{aligned}\frac{d}{dt}(f*g) &= \int_{-\infty}^{\infty} f(\tau)\frac{d}{dt}g(t-\tau)d\tau \\ &= \frac{df}{dt}*g. \end{aligned} \tag{7}$$

To compute dynamic parameters (i.e. delta and delta-delta), the differentiation operator in continuous domain can be approximated as follows:

$$\begin{aligned}\frac{d\left(\boldsymbol{y}^{\mathrm{mel}}(l)\right)}{dl} &\approx \Delta\boldsymbol{y}^{\mathrm{mel}}(l) \\ &= \frac{\sum_{\tau=-2}^{2}\tau\left(\boldsymbol{y}^{\mathrm{mel}}(l+\tau)\right)}{2\sum_{\tau=-2}^{2}\tau^2}, \end{aligned} \tag{8}$$

where l indicates the frame index. According to [9], the operator satisfies the following analogous relationship.

$$\frac{d}{dl}\left(\boldsymbol{y}^{\mathrm{mel}}(l)\right) = \boldsymbol{h}^{\mathrm{mel}}(l) * \frac{d}{dl}\left(\boldsymbol{x}^{\mathrm{mel}}(l)\right). \tag{9}$$

Therefore, the following equation is also satisfied.

$$\Delta\boldsymbol{y}^{\mathrm{mel}}(l) = \boldsymbol{h}^{\mathrm{mel}}(l) * \left(\Delta\boldsymbol{x}^{\mathrm{mel}}(l)\right). \tag{10}$$

In our model adaptation framework, Eq.(10) can be rewritten as

$$\boldsymbol{\mu}^{\mathrm{mel}}_{\Delta y}(i,m) = \alpha_{i,0}\boldsymbol{\mu}^{\mathrm{mel}}_{\Delta x}(i,m) + \sum_{j=1}^{i}\alpha_{i,i-j}\bar{\boldsymbol{\mu}}^{\mathrm{mel}}_{\Delta x}(i-j), \tag{11}$$

where $\alpha_{i,i-j}$ is the weighting coefficients given in Eq.(4). As we approximates the RIR as an exponentially decaying function, it can be rewritten by

$$\alpha_{i,i-j} = \frac{\int_{t_s(i,i-j)}^{t_e(i,i-j)}\exp\left(\frac{-3\ln(10)}{T_{60\mathrm{M}}}\right)dt}{\int_0^{\infty}\exp\left(\frac{-3\ln(10)}{T_{60\mathrm{M}}}\right)dt}, \tag{12}$$

where $T_{60\mathrm{M}}$ is the model reverberation time. In conclusion, the adaptation for both static and dynamic mean vectors can be interpreted as

$$\begin{aligned}\begin{bmatrix}\boldsymbol{\mu}^{\mathrm{mel}}_{y}(i,m)\\ \boldsymbol{\mu}^{\mathrm{mel}}_{\Delta y}(i,m)\\ \boldsymbol{\mu}^{\mathrm{mel}}_{\Delta\Delta y}(i,m)\end{bmatrix} &= \alpha_{i,0}\begin{bmatrix}\boldsymbol{\mu}^{\mathrm{mel}}_{x}(i,m)\\ \boldsymbol{\mu}^{\mathrm{mel}}_{\Delta x}(i,m)\\ \boldsymbol{\mu}^{\mathrm{mel}}_{\Delta\Delta x}(i,m)\end{bmatrix} \\ &\quad + \sum_{j=1}^{i}\alpha_{i,i-j}\begin{bmatrix}\bar{\boldsymbol{\mu}}^{\mathrm{mel}}_{x}(i-j)\\ \bar{\boldsymbol{\mu}}^{\mathrm{mel}}_{\Delta x}(i-j)\\ \bar{\boldsymbol{\mu}}^{\mathrm{mel}}_{\Delta\Delta x}(i-j)\end{bmatrix}, \end{aligned} \tag{13}$$

which is the exact same form of Eq.(5), (11).

3.2. Effects of covariance matrices

In the previous researches [6][7][8], spectral parameters are transformed between mel and log-mel domain by logarithmic and exponential operators. However, since the acoustic models in HMM framework are composed of not the sample but statistical parameters(i.e. mean, covariance), they do not follow the simple relationship.

Since the DCT operator does not change the property of Gaussian distribution, the probability density function in log-mel spectral domain can be also regarded as Gaussian distributed function. Hence, the statistical parameters in mel spectral domain can be approximated by log-normal distribution given as follows:

$$\mu^{\mathrm{mel}}(k) = \exp\left(\mu^{\log}(k) + \frac{\Sigma^{\log}(k,k)}{2}\right) \tag{14}$$

$$\Sigma^{\mathrm{mel}}(k,l) = \mu^{\mathrm{mel}}(k)\mu^{\mathrm{mel}}(l)\left(\exp(\Sigma^{\log}(k,l)) - 1\right) \tag{15}$$

$$\mu^{\log}(k) = \log(\mu^{\mathrm{mel}}(k)) - \frac{1}{2}\left(\frac{\Sigma^{\mathrm{mel}}(k,k)}{\mu^{\mathrm{mel}}(k)^2} + 1\right) \tag{16}$$

$$\Sigma^{\log}(k,l) = \log\left(\frac{\Sigma^{\mathrm{mel}}(k,l)}{\mu^{\mathrm{mel}}(k)\mu^{\mathrm{mel}}(l)} + 1\right), \tag{17}$$

where k and l is the index of feature dimension. Since the mean vectors in the log-mel spectral domain are affected by both mean vectors and covariance matrices in Eq.(14), we also needs an adaptation algorithm for updating covariance matrices. Under the assumption of independency in each state, the following equation is derived in this paper.

$$\boldsymbol{\Sigma}^{\mathrm{mel}}_{y}(i,m) = \alpha^2_{i,0}\boldsymbol{\Sigma}^{\mathrm{mel}}_{x}(i,m) + \sum_{j=1}^{i}\alpha^2_{i,i-j}\bar{\boldsymbol{\Sigma}}^{\mathrm{mel}}_{x}(i-j). \tag{18}$$

The equation is also represented by the same form as Eq.(11). Consequently, not only mean vectors but also covariance matrices can be adapted in the same framework we proposed.

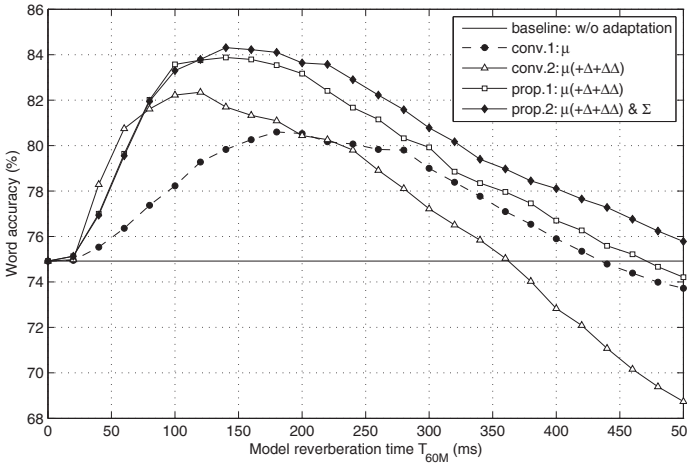

Fig. 3. Word accuracy for a variation of the model reverberation time $T_{60\mathrm{M}}$

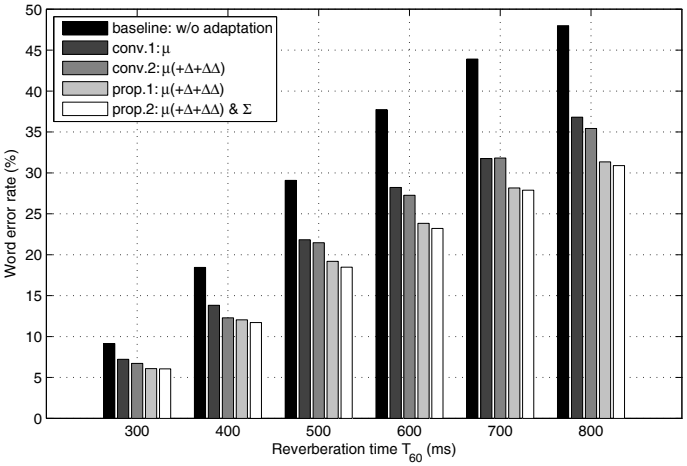

Fig. 4. Word error rate for a variation of reverberation time T_{60} in Polack's model

4. EXPERIMENTAL EVALUATION

4.1. Experimental setup

To analyze the performance of the proposed method, experiments are conducted on the TIDigits database with all utterances down-sampled at 8 kHz. To create whole word HMMs for all digits, HTK version 3.4.1 is used. Each digit model has 16 states with the mixture of three Gaussians per state, and one silence model has three states with the mixture of six Gaussians. Additionally, a short pause model for inter-word connection is created, which has a single state and shares the mixtures with the silence model. For the acoustic feature, conventional 39-dimensional mel-frequency cepstral coefficient (MFCC) vector including 0th-order coefficient is used. In the experiment, various types of RIR are used to evaluate the performance of the proposed algorithm. To obtain reverberant speech data, clean speech data are convoluted with two different types of artificial RIRs generated by Polack's model, and image method [10] [11]. Besides, measured RIRs in the real reverberant environment from Mardy database [12] are also used. In each method, two different RIRs having different reverberation time are used to verify the robustness to the estimation accuracy of reverberation time. In the experiment, the reverberation time is fixed depending on the room environment.

4.2. Experimental results

The performances of conventional and proposed methods are compared in various types of reverberant environments. In all figures and table, the *baseline* indicates the recognition system without employing any adaptation process, and *conv.1* means the algorithm that adpats only static mean vectors by conventional method given in Eq.(5). In *conv.2*, the adaptation for dynamic mean vectors are added to *conv.1* by Eq.(11). The following method *prop.1* represents the proposed algorithm that to both static and dynamic mean vectors are adapted as is given in Eq.(13). Finally, *prop.2* indicates the proposed algorithm that the adaptation of covariance matrices are also applied to *prop.1*.

Table 1. Word Accuracy (%) for connected digits recognition

Type	**Polack's**		**Image**		**Measured**	
T_{60}(ms)	500	800	540	780	598	687
baseline	74.92	55.88	78.60	63.74	76.60	82.97
conv.1	80.53	67.33	82.96	73.66	83.73	87.07
conv.2	82.32	68.25	86.92	76.63	86.25	88.18
prop.1	83.88	71.69	87.29	80.07	86.61	89.32
prop.2	84.16	72.28	87.66	80.32	86.83	89.41

In Fig. 3, the word accuracy to the different $T_{60\mathrm{M}}$ is displayed where the RIR is generated by Polack's model. The proposed method has a relative improvement of 10.41% to conventional method in terms of WER when the reverberation time is 500ms. Here, $T_{60\mathrm{M}}$ value (140ms) is much lower than the actual reverberation time, the reason for the difference is reported in [8]. In addition, the slope of right hand side of the peak point (140ms) in each curve proves that the proposed methods make the system be less sensitive to $T_{60\mathrm{M}}$ value. Fig. 4 depicts the WER obtained by varying the reverberation time of Polack's model. It shows that WER has been reduced by the proposed methods regardless of reverberation time variation. Furthermore, it shows that the amount of improvement by taking the proposed methods becomes larger as the reverberation time is longer. Finally, Table. 1 represents the word accuracy of each algorithm in six different room environments. It shows that the proposed algorithms improve

the recognition performance in the environments of both measured and artificially generated RIRs. These results confirm the superiority of the proposed methods to the conventional methods.

5. CONCLUSION

In this paper, an efficient adaptation method for the dynamic parameters of ASR system in reverberant environment has been proposed using mathematical convolution properties. The proposed method simultaneously adapts all statistical parameters in HMM within a same structure, therefore it significantly lowers computational complexity. Especially, it is more effective for the environment of long reverberation time. The proposed approach works well even the situation where the estimation accuracy of the reverberation time is imperfect.

6. REFERENCES

[1] K. Lebart, J.M. Boucher, and PN Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica united with Acustica*, vol. 87, no. 3, pp. 359–366, 2001.

[2] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *Acoustics, Speech and Signal Processing, IEEE Trans. on*, vol. 36, no. 2, pp. 145–152, 1988.

[3] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[4] M.J.F. Gales and S.J. Young, "Robust continuous speech recognition using parallel model combination," *Speech and Audio Processing, IEEE Trans. on*, vol. 4, no. 5, pp. 352–359, 1996.

[5] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM adaptation using vector Taylor series for noisy speech recognition," in *Proc. ICSLP*, 2000, vol. 3, pp. 869–872.

[6] H.G. Hirsch and H. Finster, "A new HMM adaptation approach for the case of a hands-free speech input in reverberant rooms," in *Proc. Interspeech*, 2006, pp. 781–783.

[7] H.G. Hirsch and H. Finster, "A new approach for the adaptation of HMMs to reverberation and background noise," *Speech Communication*, vol. 50, no. 3, pp. 244–263, 2008.

[8] A. Sehr, M. Gardill, and W. Kellermann, "Adapting HMMs of distant-talking ASR systems using feature-domain reverberation models," in *Proc. EUSIPCO*, 2009, pp. 540–543.

[9] RA Gopinath, MJF Gales, PS Gopalakrishnan, S. Balakrishnan Aiyer, and MA Picheny, "Robust speech recognition in noise—performance of the IBM continuous speech recogniser on the ARPA noise spoke task," in *Proceedings of the ARPA Spoken Language Systems Technology Workshop*. ARPA, 1995, pp. 127–130.

[10] J.B. Allen and D.A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am*, vol. 65, no. 4, pp. 943–950, 1979.

[11] EAP Habets, "Room impulse response (RIR) generator," *Version*, vol. 1, pp. 20080713, 2006.

[12] J.Y.C. Wen, N.D. Gaubitch, T. Myatt, and P.A. Naylor, "Evaluation of speech dereverberation algorithms using the MARDY database," *Relation*, vol. 10, no. 1.32, pp. 2640, 2008.