MAXIMUM LIKELIHOOD BASED HMM STATE FILTERING APPROACH TO MODEL ADAPTATION FOR LONG REVERBERATION

Chandra Kant Raut Takuya Nishimoto Shigeki Sagayama

Graduate School of Information Science and Technology
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656 Japan
{raut, nishi, sagayama}@hil.t.u-tokyo.ac.jp

ABSTRACT

In environment with considerably long reverberation time, each frame of speech is affected by reflected energy components from the preceding frames. Therefore to adapt parameters of a state of HMM, it becomes necessary to consider these frames, and compute their contributions to current state. However, these *clean* speech frames preceding to a state of HMM are not known during adaptation of the models. In this paper, we propose to use preceding states as units of preceding speech, and estimate their contributions to current state in maximum likelihood fashion. The experimental results on an isolated word recognition task showed significant improvement in performance of speech recognition system for reverberant speech, compared to other methods.

1. INTRODUCTION

Automatic speech recognition (ASR) systems, though usually trained with clean speech, have to operate under real-life environment for any practical purpose. One of the key factors that severely degrade the performance of speech recognition system under such practical environment is convolutional distortion caused by room reverberation or channel characteristics.

Several methods have been developed to deal with such distortion, ranging from front-end methods like inverse filtering, channel normalization and microphone-phone array based techniques to different model based approaches. Channel normalization techniques like Cepstrum Mean Subtraction (CMS) [1] and RASTA [2] have been proved effective to improve the performance of the system. Other variants of normalization technique like SNR-dependent cepstrum normalization and codeword-dependent cepstrum normalization [3] have been also proposed. Model-based approaches [4, 5] have been also applied to compensate the HMMs for channel distortion or reverberant condition. Though these methods have been proved to improve the performance of ASRs, most of them cannot perform well when reverberation time is much longer than analysis window-length and additive noise is also present. However, reverberation time longer than 100 ms is not uncommon [6] in real-life environment, e.g., in office rooms.

In our previous works [7, 8], we proposed a state splitting approach to deal with such long reverberation, which estimates preceding frames for a state of HMM and finds the compensated parameters by convolving estimated speech frames with channel parameters. In those works, channel characteristic was either assumed to be explicitly given, or estimated from adaptation data using minimum mean square error (MMSE).

This work also deals with speech distorted by long reverberation. In this work, we model the reflected component from preceding speech units in terms of preceding *states*, and estimate their contributions in maximum-likelihood manner from adaptation data. The method differs from our previous work in that it eliminates state-splitting as well as the need to have stereo-data (clean speech and its reverberated counterpart) for estimating channel parameters, making it more practical method for compensating HMMs under real-life environment.

2. EFFECT OF LONG REVERBERATION

Reverberation of a room is modeled by passing clean speech signal through a filter with impulse response of the propagation channel as transfer function, such that reverberant speech is given by

$$o[m] = h[m] * s[m] \tag{1}$$

where $s[m]$ is clean speech, $h[m]$ is impulse response of the room, $o[m]$ is reverberant speech, m is sample number and $*$ represents convolution in time domain.

Taking the short-time Fourier transform (STFT) of Eq. 1 gives

$$O(w_k, t) \approx H(w_k, t)S(w_k, t) \tag{2}$$

where t is frame number and w_k represents discrete frequency. Parameters $S(w_k, t)$, $O(w_k, t)$ and $H(w_k, t)$ are STFTs of clean speech $s[m]$, reverberant speech $o[m]$ and impulse response of channel $h[m]$, respectively. However, such a relationship is valid only when length of of $h[m]$ is shorter than window-size used for STFT.

The effect of reverberation with reverberation time (T_{60}) longer than the analysis window size, on short-time Fourier transform of speech is usually approximated by

$$O(w_k, t) \approx H(w_k, t) * S(w_k, t). \tag{3}$$

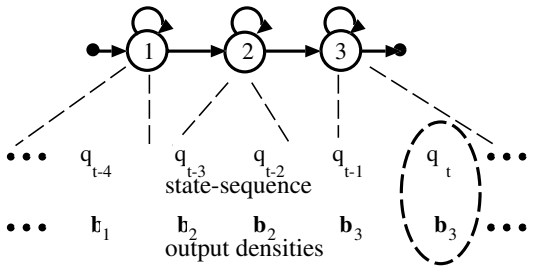

Fig. 1. HMM Adaptation: To compensate state output $\boldsymbol{b}_j$ at state $q_t = j$ for long reverberation, by accounting energy components from previous frames, knowledge of clean observations at time $t-1$, $t-2$ and so on is necessary.

In other words, the effect of long reverberation is no more multiplicative in linear spectral domain, rather it is convolutional.

3. ACCOUNTING CONVOLUTIONAL DISTORTION

Eq. 3 shows that the spectral parameters of reverberant speech at frame t do not depend only upon this frame, but also upon the preceding frames at $t-1$, $t-2$ and so on. Therefore, to adapt the output distribution $\boldsymbol{b}_j$ at state $q_t = j$ of given HMM [Fig. 1], frames occured at time $t-1$, $t-2$ and so on should be considered. However, with such a conventional HMM used in most of speech recognition systems, nothing can be inferred deterministically about the observations, and even the state sequence preceding to a given state cannot be known (as adaptation under consideration is *not* decoding-time adaptation and therefore does not use observations, and in such case even most likely state sequence preceding to a state cannot be estimated).

In [7, 8], we proposed a state splitting approach to estimate preceding frames for a given state of HMM by using state's duration information to estimate state-sequence and using composite mean of the output distributions of those states as preceding observation sequence. Such estimation of preceding frames becomes necessary, when the channel parameters are estimated for the frame-level convolution. Assuming that *average occupation of states does not vary significantly*, or *if each state is usually occupied only once*, we approximate the convolutional distortion by filtering of *states* (convolution over states) directly, with some optimal channel parameters, without necessitating to work at frame-level and estimate preceding *frames*. Such convolution over states, in linear spectral domain, is represented as

$$\begin{aligned}\hat{\boldsymbol{O}}(j) &= \boldsymbol{\alpha}_0\boldsymbol{S}(j) + \boldsymbol{\alpha}_1\boldsymbol{S}(j-1) + \boldsymbol{\alpha}_2\boldsymbol{S}(j-2) \\ &\quad + \ldots + \boldsymbol{\alpha}_{N-1}\boldsymbol{S}(j-N+1) \qquad (4)\end{aligned}$$

where $\boldsymbol{S}$ and $\boldsymbol{\alpha}_i$ are clean speech state output distribution and state-filter coefficients, respectively (separate equations in terms of mean and covariance matrix defined later in Eq. 11 and 12); k represents the dimension of the parameter, and j represents HMM states of model (please see Fig. 2 for interpretation of j). Left contexts of models can be used to account the effect of preceding models; in their absence, only the available preceding states of current model can be used.

With this approach, need for state-splitting and estimation of preceding frames using state's duration information is eliminated; however, optimal $\boldsymbol{\alpha}_i$ should be estimated for such convolution over states. The next section describes its estimation by maximum likelihood approach.

4. MAXIMUM-LIKELIHOOD ESTIMATION OF STATE-FILTER COEFFICIENTS

Reverberant speech model λ_O is composed by using clean speech model λ_S and state-level channel parameters $\boldsymbol{A} = \{\boldsymbol{\alpha}_0, \ldots, \boldsymbol{\alpha}_i, \ldots, \boldsymbol{\alpha}_{N-1}\}$. Parameter α_{ik} (k : dimension of speech parameter) is estimated by maximizing Viterbi-likelihood score $P(\boldsymbol{O}, \boldsymbol{q}|\boldsymbol{A}, \lambda_S)$ or $P(\boldsymbol{O}, \boldsymbol{q}|\lambda_O)$ of training observation $\boldsymbol{O} = \{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T\}$ over most likelihood state sequence $\boldsymbol{q} = \{q_1, \ldots, q_T\}$ given by Viterbi algorithm, as

$$\hat{\alpha}_{ik} = \arg\max_{\alpha_{ik}} P(\boldsymbol{O}|\boldsymbol{\alpha}_{0,\ldots,N-1}, \lambda_S). \qquad (5)$$

Maximization of $P(\boldsymbol{O}, \boldsymbol{q}|\boldsymbol{A}, \lambda_S)$ is done in iterative manner by steepest-descent method, by defining new estimate of α_{ik} at pth iteration as

$$\begin{aligned}\alpha_{ik}(p) &= \alpha_{ik}(p-1) \\ &\quad + \epsilon \frac{\partial \log\left(P(\boldsymbol{O}|\boldsymbol{\alpha}_{0,\ldots,N-1}, \lambda_S)\right)}{\partial \alpha_{ik}} \qquad (6)\end{aligned}$$

where ϵ is gradient scaling factor.

Computation of likelihood $P(\boldsymbol{O}, \boldsymbol{q}|\boldsymbol{A}, \lambda_S)$ (and its maximization) is done using cepstrum domain parameters, whereas composition of reverberant speech model λ_O by using λ_S, $\boldsymbol{A}$ and additive noise (when considered) is done in linear-spectral domain, under PMC framework [4]. Therefore, such estimation involves conversion of parameters between these domains at each iteration. The parameters in cepstrum, log and linear spectral domains are specified by subscript cep, lg and lin, respectively.

Transformation of models from cepstrum-domain to log spectral domain is done as

$$\boldsymbol{\mu}^{S_{\text{lg}}} = \boldsymbol{C}^{-1}\boldsymbol{\mu}^{S_{\text{cep}}} \qquad (7)$$

$$\boldsymbol{\Sigma}^{S_{\text{lg}}} = \boldsymbol{C}^{-1}\boldsymbol{\Sigma}^{S_{\text{cep}}}(\boldsymbol{C}^{-1})^T \qquad (8)$$

where $\boldsymbol{C}$ is the discrete Cosine transform (DCT) matrix. These parameters in log spectral domain are further transformed to linear spectral domain, by using

$$\mu_k^{S_{\text{lin}}} = \exp\left(\mu_k^{S_{\text{lg}}} + \frac{\Sigma_{kk}^{S_{\text{lg}}}}{2}\right) \qquad (9)$$

$$\Sigma_{kl}^{S_{\text{lin}}} = \mu_k^{S_{\text{lin}}}\mu_l^{S_{\text{lin}}}\left(\exp(\Sigma_{kl}^{S_{\text{lg}}}) - 1\right) \qquad (10)$$

where k and l are parameter indices.

In linear spectral domain, model for reverberant speech is composed by using clean speech model and estimated α_{ik} as

$$\begin{aligned}\mu_k^{O_{\rm lin}}(j) &= \alpha_{0k}\mu_k^{S_{\rm lin}}(j) \\ &\quad + \alpha_{1k}\bar{\mu}_k^{S_{\rm lin}}(j-1) + \alpha_{2k}\bar{\mu}_k^{S_{\rm lin}}(j-2) \\ &\quad + \ldots + \alpha_{N-1,k}\bar{\mu}_k^{S_{\rm lin}}(j-N+1) \quad (11)\\ \Sigma_{kl}^{O_{\rm lin}}(j) &= \alpha_{0k}\alpha_{0l}\Sigma_{kl}^{S_{\rm lin}}(j) \quad (12)\end{aligned}$$

From the preceding states, only composite mean (distinguished by overbar) from single component distribution corresponding to Gaussian mixture model of output distributions are used. Such single component composite distribution from M-mixture GMM can be obtained, as

$$\bar{\boldsymbol{\mu}}_{\rm cep} = \sum_{m=1}^{M} c_m \boldsymbol{\mu}_{m,\rm cep} \quad (13)$$

$$\bar{\boldsymbol{\Sigma}}_{\rm cep} = \sum_{m=1}^{M} c_m \left(\boldsymbol{\Sigma}_{m,\rm cep} + \boldsymbol{\mu}_{m,\rm cep}\boldsymbol{\mu}_{m,\rm cep}^T\right) - \bar{\boldsymbol{\mu}}_{\rm cep}\bar{\boldsymbol{\mu}}_{\rm cep}^T \quad (14)$$

where m represents mixture component, and c_m is mixture weight. Furthermore, effect of preceding states on covariance matrix is ignored.

Once the models are adapted in linear spectral domain, they are transformed back to log spectral domain by using

$$\begin{aligned}\mu_k^{O_{\rm lg}} &= \log(\mu_k^{O_{\rm lin}}) - \frac{1}{2}\log\left(\frac{\Sigma_{kk}^{O_{\rm lin}}}{{\mu_k^{O_{\rm lin}}}^2} + 1\right) \quad (15)\\ \Sigma_{kl}^{O_{\rm lg}} &= \log\left(\frac{\Sigma_{kl}^{O_{\rm lin}}}{\mu_k^{O_{\rm lin}}\mu_l^{O_{\rm lin}}} + 1\right), \quad (16)\end{aligned}$$

and to cepstrum domain by using

$$\begin{aligned}\boldsymbol{\mu}^{O_{\rm cep}} &= \boldsymbol{C}\boldsymbol{\mu}^{O_{\rm lg}} \quad (17)\\ \boldsymbol{\Sigma}^{O_{\rm cep}} &= \boldsymbol{C}\boldsymbol{\Sigma}^{O_{\rm lg}}\boldsymbol{C}^T. \quad (18)\end{aligned}$$

Such formulations for transformation of parameters from one-domain to another and composition of models are used while estimating α_{ik} as well. We use similar approach as in [5] to maximize likelihood and estimate filter coefficients. As under large mixture GMMs, estimation of α_{ik} becomes complex, they can be first reduced to single-component distribution using Eqs.13 and 14 and used while estimating α_{ik}. The new estimate for α_{ik}, under single-mixture case, is given by

$$\begin{aligned}\alpha_{ik}(p) &= \alpha_{ik}(p-1) \\ &\quad + \epsilon\frac{\partial}{\partial\alpha_{ik}}\sum_{\forall t}\{-\frac{1}{2}\log\left((2\pi)^D \mid \boldsymbol{\Sigma}_t^{O_{\rm cep}} \mid\right) \\ &\quad - \frac{1}{2}(\boldsymbol{O}_t - \boldsymbol{\mu}_t^{O_{\rm cep}})^T\boldsymbol{\Sigma}_t^{O_{\rm cep}^{-1}}(\boldsymbol{O}_t - \boldsymbol{\mu}_t^{O_{\rm cep}})\} \quad (19)\end{aligned}$$

where $\{(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \ldots, (\boldsymbol{\mu}_T, \boldsymbol{\Sigma}_T)\}$ corresponds to output distributions of most likely state-sequence decoded by Viterbi algorithm. Ignoring the change in covariance w.r.t. α_{ik} leads to

$$\begin{aligned}\alpha_{ik}(p) &= \alpha_{ik}(p-1) \\ &\quad + \epsilon\sum_{\forall t}\left(\frac{1}{2}\frac{\partial{\boldsymbol{\mu}_t^{O_{\rm cep}}}^T}{\partial\alpha_{ik}}\boldsymbol{\Sigma}_t^{O_{\rm cep}^{-1}}(\boldsymbol{O}_t - \boldsymbol{\mu}_t^{O_{\rm cep}})\right. \\ &\quad \left. + \frac{1}{2}(\boldsymbol{O}_t - \boldsymbol{\mu}_t^{O_{\rm cep}})^T\boldsymbol{\Sigma}_t^{O_{\rm cep}^{-1}}\frac{\partial\boldsymbol{\mu}_t^{O_{\rm cep}}}{\partial\alpha_{ik}}\right) \quad (20)\\ &= \alpha_{ik}(p-1) \\ &\quad + \epsilon\sum_{\forall t}\left(\frac{1}{2}(\boldsymbol{C}\frac{\partial\boldsymbol{\mu}_t^{O_{\rm lg}}}{\partial\alpha_{ik}})^T\boldsymbol{\Sigma}_t^{O_{\rm cep}^{-1}}(\boldsymbol{O}_t - \boldsymbol{\mu}_t^{O_{\rm cep}})\right. \\ &\quad \left. + \frac{1}{2}(\boldsymbol{O}_t - \boldsymbol{\mu}_t^{O_{\rm cep}})^T\boldsymbol{\Sigma}_t^{O_{\rm cep}^{-1}}\boldsymbol{C}\frac{\partial\boldsymbol{\mu}_t^{O_{\rm lg}}}{\partial\alpha_{ik}}\right). \quad (21)\end{aligned}$$

The term $\partial\boldsymbol{\mu}_t^{O_{\rm lg}}/\partial\alpha_{ik}$ (each kth component represented as $\partial\mu_k^{O_{\rm lg}}(j)/\partial\alpha_{ik}$, where j is the aligned state to frame t of training observation) can be obtained by taking derivative of Eq. 15 as

$$\begin{aligned}\frac{\partial\mu_k^{O_{\rm lg}}(j)}{\partial\alpha_{ik}} &= \frac{\mu_k^{S_{\rm lin}}(j-i)}{\mu_k^{O_{\rm lin}}(j)} \\ &\quad + \frac{\Sigma_{kk}^{O_{\rm lin}}(j)\mu^{S_{\rm lin}}(j-i)}{\mu_k^{O_{\rm lin}}(j)\Sigma_{kk}^{O_{\rm lin}}(j) + \left(\mu_k^{O_{\rm lin}}(j)\right)^3}. \quad (22)\end{aligned}$$

While also considering additive noise, mean and covariance matrix terms for it can be included in Eqs. 11 and 12, and can be estimated together, or if already estimated (e.g. using signal during non-speech activity), they can be used during estimation of α_{ik}. When likelihood score converges for the training observation, the estimated values of α_{ik} are used to transform all the clean models to models for reverberant speech. The procedure is also depicted in Fig. 2.

5. EVALUATION

The proposed method was evaluated on a speaker-dependent isolated word recognition task. Clean speech HMMs were trained with 2620 words of the same speaker taken from ATR speech database A-Set. Clean speech HMMs comprised of 425 context-dependent biphone models with left-context, each with three emitting states single mixture Gaussian model. The speech signal was single channel with sampling frequency of 16 kHz. The speech signal was analyzed with Hamming window of 25 ms window-size and frame shift of 10 ms into 13-dimensional MFCC feature vectors including 0th-order coefficient, using 24 mel filter-banks. The test set consisted of 655 words of the same speaker taken exclusively from the ATR speech database A-set, and HTK 3.1 was used as decoder.

For evaluation, reverberant speech was simulated by a linear convolution of clean speech and impulse responses of the environment (viz. E1B and OFC) taken from RWCP Sound Scene Database in Real Environment. The performance of the speech recognition system decreased with clean

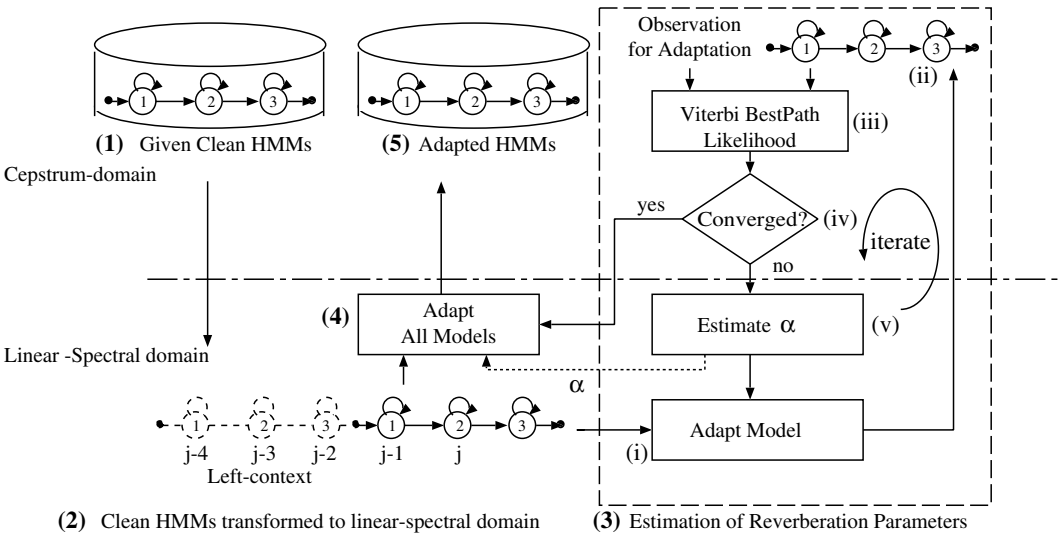

Fig. 2. Adaptation of model parameters

Table 1. Experimental Results (Word Recognition Rate %)

Data	T_{60}	Clean	CMS	SF(N=4)
Clean	–	97.9	–	–
E1B	310 ms	67.6	77.3	83.2
OFC	780 ms	44.8	47.5	72.5

model as listed under "Clean" in Table 1 for the reverberant speech. The recognition performance of reverberant speech was evaluated with CMS as well. For this purpose, CMS was performed on the same training set data, and the model was retrained with it. CMS was applied to test set also, and performance was evaluated with the retrained model.

To evaluate the proposed method, ten words of reverberant speech was used as adaptation data to estimate α_{ik}, with filter-order of $N = 4$, and states of left-context were considered for the convolution.

Experimental results as listed under Table 1 show better performance of proposed state-filtering (SF) approach compared to clean model and CMS, which demonstrates its effectiveness for recognizing reverberant speech. With longer reverberation time, the improvement with state-filtering approach is more pronounced compared to other methods.

6. CONCLUSION

In this paper, we proposed a method for model adaptation based on state-filtering, for reverberant speech with considerably long reverberation time. The filter coefficients are estimated by using maximum-likelihood approach, using small amount of training data. The experimental result shows the effectiveness of the method for improving performance of the system under reverberant condition.

Future work includes evaluation of the method on large vocabulary continuous speech recognition task and with Gaussian mixture models. Its application for adapting models for moving-speaker case will be also investigated.

7. REFERENCES

[1] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Am.*, vol. 55, pp. 1304–1312, 1974.

[2] B. E. Kingsbury and N. Morgan, "Recognizing reverberant speech with RASTA-PLP," in *Proc. ICASSP*, Munich, Germany, 1997, pp. 1259–1262.

[3] A. Acero and R. M. Stern, "Environmental robustness in automatic speech recognition," *Proc. ICASSP*, pp. 849–852, 1990.

[4] M. J. F. Gales and S. J. Young, "Robust speech recognition in additive and convolutional noise using parallel model combination," *Computer Speech and Language*, vol. 9, pp. 289–307, 1995.

[5] Y. Minami and S. Furui, "A maximum likelihood procedure for a universal adaptation method based on HMM composition," in *Proc. ICASSP*, 1995, pp. 129– 132.

[6] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall PTR, New Jersey, 1st edition, 2001.

[7] C. K. Raut, T. Nishimoto, and S. Sagayama, "Model adaptation by state splitting of HMM for long reverberation," in *Proc. Interspeech*, Sep. 2005.

[8] C. K. Raut, T. Nishimoto, and S. Sagayama, "Model convolution by state splitting of HMM for robust speech recognition in presence of convolutional noise," in *Proc. The Acoustical Society of Japan*, Mar. 2005, vol. 3-5-5, pp. 85–86.