Maximum Likelihood Based Compensation of HMM Parameters for Channel Distortion *

◎ Chandra Kant Raut Takuya Nishimoto Shigeki Sagayama (The University of Tokyo)

1 Introduction

The performance of automatic speech recognition (ASR) systems trained with clean speech degrades severely due to channel distortion. Several methods have been developed to deal with such distortion, ranging from front-end methods like inverse filtering, channel normalization techniques like Cepstrum Mean Subtraction (CMS) [1], and microphone-phone array based methods to different model based approaches [2, 3]. 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 in real-life environment, e.g., in office rooms.

This work deals with long channel distortion (reverberation time T_{60} longer than analysis window length), and compensates the model parameters to reduce the mismatch between model and distorted speech. In this work, we model energy component contributed by preceding speech units in terms of preceding *states*, and estimate their contributions in maximum-likelihood manner from adaptation data.

2 Effect of Channel Distortion and Model Adaptation

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

$$O(w_i, t) \approx H(w_i, t) * S(w_i, t) \tag{1}$$

where t is frame number, w_i is discrete frequency and * represents convolution along frame. Parameters $S(w_i, t)$, $H(w_i, t)$ and $O(w_i, t)$ are STFTs of clean speech $s[m]$, impulse response $h[m]$ characterizing channel distortion, and distorted speech $o[m] = h[m] * s[m]$, respectively.

Therefore, the effect of long reverberation is no more multiplicative in spectral domain, rather it is convolutional. Eq. 1 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. Considering the effect from model-domain perspective, to adapt the parameters of state j of given HMM [Fig. 1] to model the distorted speech, preceding segments (frames) of speech need to be considered. As such clean speech frames preceding to a state are not known during adaptation, we use mean vectors of preceding states as segments of preceding speech, and approximate the distorted speech by filtering of *states* directly, with some optimal channel parameters.

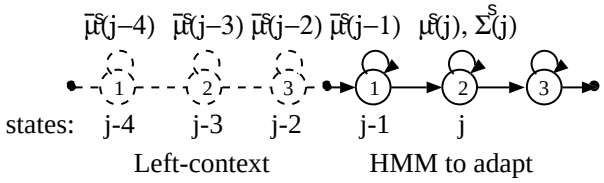

Figure 1: Model Adaptation: Parameters of preceding states used for accounting effect of preceding speech segments caused by long reverberation time.

After transforming the cepstrum-domain model parameters to linear spectral domain, the model parameters of a state, viz. mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, for distorted speech are estimated as:

$$\begin{aligned}\mu_k^{O_{\text{lin}}}(j) &= \alpha_{0k}\mu_k^{S_{\text{lin}}}(j) \\ &\quad + \alpha_{1k}\bar{\mu}_k^{S_{\text{lin}}}(j-1) + \alpha_{2k}\bar{\mu}_k^{S_{\text{lin}}}(j-2) \\ &\quad + \ldots + \alpha_{N-1,k}\bar{\mu}_k^{S_{\text{lin}}}(j-N+1) \qquad (2)\\ \Sigma_{kl}^{O_{\text{lin}}}(j) &= \alpha_{0k}\alpha_{0l}\Sigma_{kl}^{S_{\text{lin}}}(j) \qquad (3)\end{aligned}$$

where j represents state number of HMM, k, l represent dimensions of parameters, α_{ik}s are filter coefficients, and subscript *lin* signifies linear domain parameters. From the preceding states, only the composite mean (distinguished by overbar) from single component distribution corresponding to Gaussian mixture model of output distributions are used. Left contexts of models can be used to account the effect of states from preceding models.

Once the models are adapted in linear spectral domain, they are transformed back to log spectral domain (represented by subscript lg) by using

$$\mu_k^{O_{\text{lg}}} = \log(\mu_k^{O_{\text{lin}}}) - \frac{1}{2}\log\left(\frac{\Sigma_{kk}^{O_{\text{lin}}}}{{\mu_k^{O_{\text{lin}}}}^2} + 1\right) \tag{4}$$

$$\Sigma_{kl}^{O_{\text{lg}}} = \log\left(\frac{\Sigma_{kl}^{O_{\text{lin}}}}{\mu_k^{O_{\text{lin}}}\mu_l^{O_{\text{lin}}}} + 1\right), \tag{5}$$

and to cepstrum domain by using $\boldsymbol{\mu}^{O_{\text{cep}}} = \boldsymbol{C}\boldsymbol{\mu}^{O_{\text{lg}}}$ and $\boldsymbol{\Sigma}^{O_{\text{cep}}} = \boldsymbol{C}\boldsymbol{\Sigma}^{O_{\text{lg}}}\boldsymbol{C}^T$, where $\boldsymbol{C}$ is discrete Cosine transform (DCT) matrix.

The optimal value of coefficients α_{ik} representing degree of contribution of preceding and current states are estimated by maximizing likelihood of channel-distorted adaptation data.

*チャネル歪みのための最尤法による HMM パラメータ適応, チャンドラ カント ラウト, 西本 卓也, 嵯峨山 茂樹 (東大・情報理工)

3 Maximum-Likelihood Estimation of State Filter coefficients

The model λ_O for distorted speech is composed by using clean speech model λ_S and estimated parameters $\boldsymbol{A} = \{\boldsymbol{\alpha}_0, \ldots, \boldsymbol{\alpha}_{N-1}\}$. The parameters α_{ik} are 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

$$\hat{\alpha}_{ik} = \arg\max_{\alpha_{ik}} P(\boldsymbol{O}|\boldsymbol{\alpha}_{0,\ldots,N-1}, \lambda_S). \quad (6)$$

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

$$\alpha_{ik}(p) = \alpha_{ik}(p-1) + \epsilon \frac{\partial \log\left(P(\boldsymbol{O}|\boldsymbol{\alpha}_{0,\ldots,N-1}, \lambda_S)\right)}{\partial \alpha_{ik}} \quad (7)$$

where ϵ is scaling factor.

We use similar approach as in [3] to maximize likelihood and estimate filter coefficients. As estimation of α_{ik} becomes complex for GMMs with large mixture components, they can be first reduced to single-mixture components. The new estimate for α_{ik}, under single-mixture case, is given by

$$\begin{aligned}\alpha_{ik}(p) = \alpha_{ik}(p-1) &+ \epsilon \frac{\partial}{\partial \alpha_{ik}} \sum_{\forall t} \{-\frac{1}{2} \log\left((2\pi)^D \mid \boldsymbol{\Sigma}_t^{O_{\text{cep}}} \mid\right) \\ &- \frac{1}{2} (\boldsymbol{O}_t - \boldsymbol{\mu}_t^{O_{\text{cep}}})^T \boldsymbol{\Sigma}_t^{O_{\text{cep}}^{-1}} (\boldsymbol{O}_t - \boldsymbol{\mu}_t^{O_{\text{cep}}})\} \quad (8)\end{aligned}$$

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

$$\begin{aligned}\alpha_{ik}(p) &= \alpha_{ik}(p-1) \\ &\quad + \epsilon \sum_{\forall t} \left(\frac{1}{2} \frac{\partial \boldsymbol{\mu}_t^{O_{\text{cep}}{}^T}}{\partial \alpha_{ik}} \boldsymbol{\Sigma}_t^{O_{\text{cep}}^{-1}} (\boldsymbol{O}_t - \boldsymbol{\mu}_t^{O_{\text{cep}}}) \right. \\ &\quad \left. + \frac{1}{2} (\boldsymbol{O}_t - \boldsymbol{\mu}_t^{O_{\text{cep}}})^T \boldsymbol{\Sigma}_t^{O_{\text{cep}}^{-1}} \frac{\partial \boldsymbol{\mu}_t^{O_{\text{cep}}}}{\partial \alpha_{ik}} \right) \\ &= \alpha_{ik}(p-1) \\ &\quad + \epsilon \sum_{\forall t} \left(\frac{1}{2} (\boldsymbol{C} \frac{\partial \boldsymbol{\mu}_t^{O_{\text{lg}}}}{\partial \alpha_{ik}})^T \boldsymbol{\Sigma}_t^{O_{\text{cep}}^{-1}} (\boldsymbol{O}_t - \boldsymbol{\mu}_t^{O_{\text{cep}}}) \right. \\ &\quad \left. + \frac{1}{2} (\boldsymbol{O}_t - \boldsymbol{\mu}_t^{O_{\text{cep}}})^T \boldsymbol{\Sigma}_t^{O_{\text{cep}}^{-1}} \boldsymbol{C} \frac{\partial \boldsymbol{\mu}_t^{O_{\text{lg}}}}{\partial \alpha_{ik}} \right). \quad (9)\end{aligned}$$

The term $\partial \boldsymbol{\mu}_t^{O_{\text{lg}}} / \partial \alpha_{ik}$ (each kth component represented as $\partial \mu_k^{O_{\text{lg}}}(j) / \partial \alpha_{ik}$, where j is the aligned state to frame t of training observation) can be obtained by taking derivative of Eq. 4 as

Table 1: Expt. Result (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

$$\begin{aligned}\frac{\partial \mu_k^{O_{\text{lg}}}(j)}{\partial \alpha_{ik}} &= \frac{\mu_k^{S_{\text{lin}}}(j-i)}{\mu_k^{O_{\text{lin}}}(j)} \\ &\quad + \frac{\Sigma_{kk}^{O_{\text{lin}}}(j) \mu^{S_{\text{lin}}}(j-i)}{\mu_k^{O_{\text{lin}}}(j) \Sigma_{kk}^{O_{\text{lin}}}(j) + \left(\mu_k^{O_{\text{lin}}}(j)\right)^3}. \quad (10)\end{aligned}$$

4 Evaluation and Conclusion

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. Single-channel speech signal, sampled at 16 kHz, was analyzed with Hamming window of 25 ms frame length and frame shift of 10 ms into 13-dimensional MFCC feature vectors including 0th-order coefficient, using 24 mel filter-banks.

For evaluation, reverberant speech was simulated by a linear convolution of clean speech and impulse responses (viz. E1B and OFC) taken from RWCP Sound Scene Database in Real Environment. 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 proposed method, ten words of distorted speech was used as adaptation data to estimate α_{ik}, with filter-order of $N = 4$, and states of left-contexts were considered as well for the adaptation.

The experimental result as listed under Table 1 shows better performance of proposed state-filtering (SF) approach compared to clean model and CMS, which demonstrates its effectiveness to improve the performance of the speech recognition system for channel-distorted speech.

Future work includes evaluation of the method on large vocabulary continuous speech recognition task and with Gaussian mixture models.

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] 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.

[3] 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.