MODEL ADAPTATION FOR LONG CONVOLUTIONAL DISTORTION BY MAXIMUM LIKELIHOOD BASED STATE FILTERING APPROACH

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 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 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 segments, estimate their contributions to current state in maximum likelihood manner, and adapt models by accounting their contributions. When clean models were adapted by proposed method for a speaker-dependent isolated word recognition task, word accuracy of the system typically increased from 67.6% to 83.2%, and from 44.8% to 72.5%, for channel distorted speech simulated by linear convolution of clean speech and impulse responses with reverberation time (T_{60}) of 310 ms and 780 ms, respectively.

1. INTRODUCTION

Automatic speech recognition (ASR) systems, though usually trained with clean speech, have to operate under real-life condition, which includes hands-free communication (with far-field microphone), reverberant rooms, and telephone networks. Convolutional distortion caused by channel characteristics or reverberation in such environment can severely degrade the performance of the system, and can make it completely useless for any practical purpose. Therefore, any practically usable system must be able to cope with such convolutional distortion present in speech. A number of techniques have been developed over years to deal with such distortion, ranging from front-end methods to several model-based approaches.

Front-end methods include inherently robust and enhancement based techniques. Inherently robust techniques focus on distortion-resistant features and distance measures. Channel normalization techniques like Cepstrum Mean Subtraction (CMS) [1] and RASTA [2] have been proved effective to improve the performance of the system. Time derivatives of cepstra [3] have been very effective to improve recognition rate under noisy and reverberant condition, as well as under clean environment. Probabilistic signal bias estimation and removal technique [4], which iteratively estimates bias by maximizing likelihood of speech model, was also proposed. Enhancement-based techniques, on the other side, essentially attempt to transform distorted speech parameters into clean speech parameters. Such techniques include inverse filtering and various microphone-array based techniques. Codeword-dependent cepstrum normalization (CDCN) [5] has been also proved effective for recognition of speech from far-field microphone.

Model-based approaches, on the other hand, attempt to transform clean speech models to distorted ones, thus reducing the mismatch between training and testing environment. Parallel model combination (PMC) [6], though mostly popular for compensating additive noise, has been used to compensate for convolutional distortion as well, by estimating convolutional noise component from adaptation data and generalized speech model. Similarly, universal adaptation method [7] compensates for channel distortion along with additive noise by maximizing likelihood of adaptation data in log-normal domain. Compensation to convolutional distortion has been considered in vector Taylor series (VTS) [8] and other polynomial based approaches as well. Other general model based techniques that have been proved effective for adapting models for channel distorted speech include maximum likelihood linear regression (MLLR) [9] and maximum a posteriori (MAP) [10] estimation.

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. However, reverberation time longer than 100 ms is not uncommon [11] in real-life environment, e.g., in office rooms. Feature-based methods like inverse-filtering and RASTA-PLP, despite their own limitations, have been more considerate of long convolutional distortion than most of the popular and successful model based approaches like PMC, VTS, universal adaptation and others, where compensation to convolutional distortion is generally based on a single state, without explicitly considering effect from preceding states or speech segments. Some of the model based approaches that explicitly consider the effect of preceding speech segments for adaptation include first-order linear prediction [12] and our previous work [13] based on state-splitting. In [12], energy component from preceding frames is estimated by first order linear prediction from (single) last frame of observation, and models are adapted at *each* frame, which is computationally very expensive and inefficient. In our previous work [13], we proposed a state splitting approach to estimate preceding frames for a given state of HMM, which are used to compensate parameters of the state by convolving with channel parameters. However, the method needs stereo data for estimating channel parameters. Besides, model structure is changed and large number of states are introduced in the system making it quite complex and inefficient for decoding.

This work is also a model-based approach and it considers adaptation for long channel distortion, when reverberation time (T_{60}) is much longer than analysis window length. In this work, we model energy component contributed by preceding speech segments in terms of preceding *states*, and estimate amount of their contributions in maximum-likelihood manner from adaptation data. Only few seconds of distorted

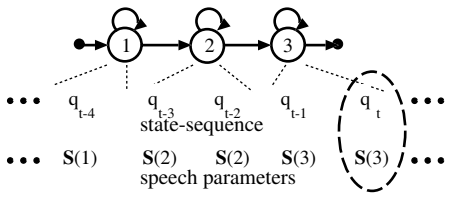

Fig. 1. HMM Adaptation: To compensate state output $\boldsymbol{S}(j)$ at state $q_t = j$ for long convolutional distortion, knowledge of *clean observations* at frames $t-1$, $t-2$ and so on is necessary.

speech is required without demanding stereo recordings; and besides that, the method is flexible for simultaneously compensating additive noise as well.

2. EFFECT OF LONG CONVOLUTIONAL DISTORTION

Convolutional distortion in speech is modeled as filtering of clean speech signal $s[m]$ by channel characteristics (impulse response) $h[m]$, such that distorted speech $o[m]$ in time-domain is given as $h[m] * s[m]$ (where m is sample number, and $*$ represents convolution in time domain).

When reverberation time T_{60} of $h[m]$ is much shorter than window-size used for short-time Fourier transform (STFT), STFT of distorted speech is given as

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

where t is frame number and w_k represents discrete frequency. Parameters $S(w_k, t)$ and $H(w_k, t)$ are STFTs of clean speech $s[m]$ and impulse response of channel $h[m]$, respectively.

However, when reverberation time T_{60} is much longer than window-size, such a relationship is no more valid, and STFT of distorted speech is usually approximated by

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

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

3. ACCOUNTING CONVOLUTIONAL DISTORTION

Eq. 2 can be rewritten by limiting convolution up to finite length L, as

$$\begin{aligned} O(w_k, t) &\approx H(w_k, 0)S(w_k, t) + H(w_k, 1)S(w_k, t-1) \\ &+ \ldots + H(w_k, L-1)S(w_k, t-L+1). \end{aligned} \tag{3}$$

Eq. 3 shows that the spectral parameters of distorted speech at frame t do not depend only upon this frame, but also upon the preceding *clean* frames at $t-1$, $t-2$ and so on. From model-domain perspective, this implies that compensated output distribution $\hat{\boldsymbol{O}}(j)$ at state $q_t = j$ of given HMM [Fig. 1] depends upon the *clean* observations at frames $t-1$,$t-2$ and so on, which are not known in any case (whether front-end methods or model-based approach).

Besides, model adaptation being considered is *not* decoding-time, which means that models are adapted once, without even using incoming distorted observations, and used until channel characteristics changes significantly. In such case, nothing can be inferred deterministically about the frames of clean observations that will precede to a given state of HMM, and even the most likely state occupations cannot be known.

However, states (but not their exact occupations) preceding to a given state can be known up to some extent, from the structure of models and given contexts (like triphones or biphones). For example, while adapting state 2 of HMM in Fig. 2, we know that state 1 of the model will precede it, and beyond that states of its left context will occur. Therefore, we propose to model the frame-level convolutional distortion as given in Eq.3, directly in terms of these preceding states, such that distorted speech parameters at state j are represented by filtering of states (or state-level convolution), as

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

where $\boldsymbol{S}(j)$ is parameter of state j of clean speech models (please see Fig. 2 for interpretation of j) and $\boldsymbol{\alpha}_i$ is state-level filter coefficient (cf. frame-level filter coefficients $H(w_k, t)$ of Eq. 3). Separate equations in terms of mean and covariance matrix are defined later in Eq. 11 and 12. 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. Eq. 4 with only first term on the right-hand side and parameters representing means can be seen as similar to MLLR mean compensation (without bias term).

Representing convolutional distortion directly in terms of preceding states can be regarded as considering effect of preceding *blocks of frames* (represented by each state) on current state rather than of individual frames. As such 'blocks' will be of variable size, estimation of optimal values of state-level filter coefficients are important so that they can be applied over different states. The next section describes their estimation from few seconds of distorted speech data by maximum likelihood approach.

4. MAXIMUM-LIKELIHOOD ESTIMATION OF STATE-FILTER COEFFICIENTS

Corrupted speech model λ_O is composed by using state-level channel parameters $\boldsymbol{A} = \{\boldsymbol{\alpha}_0, \ldots, \boldsymbol{\alpha}_i, \ldots, \boldsymbol{\alpha}_{N-1}\}$ and clean speech models λ_S. 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 adaptation observation $\boldsymbol{O} = \{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T\}$ over most likely state sequence $\boldsymbol{q} = \{q_1, \ldots, q_T\}$ given by Viterbi algorithm, as

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

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

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

where ϵ is 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 corrupted speech model λ_O by using λ_S, $\boldsymbol{A}$ and additive noise (when considered) is done in linear-spectral domain, under PMC framework [6]. 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

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

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

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

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

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

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

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

The covariance matrix can be retained unchanged as in Eq. 12. Also, for adaptation of means, only composite mean (shown by overbar) of preceding states from single component distribution corresponding to Gaussian mixture model of their 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.

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

$$\mu_k^{O_{\rm lg}} = \log(\mu_k^{O_{\rm lin}}) - \frac{1}{2}\log\left(\frac{\Sigma_{kk}^{O_{\rm lin}}}{{\mu_k^{O_{\rm lin}}}^2} + 1\right) \quad (15)$$

$$\Sigma_{kl}^{O_{\rm lg}} = \log\left(\frac{\Sigma_{kl}^{O_{\rm lin}}}{\mu_k^{O_{\rm lin}}\mu_l^{O_{\rm lin}}} + 1\right), \quad (16)$$

and to cepstrum domain by using

$$\boldsymbol{\mu}^{O_{\rm cep}} = \boldsymbol{C}\boldsymbol{\mu}^{O_{\rm lg}} \quad (17)$$

$$\boldsymbol{\Sigma}^{O_{\rm cep}} = \boldsymbol{C}\boldsymbol{\Sigma}^{O_{\rm lg}}\boldsymbol{C}^T. \quad (18)$$

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 [7] to maximize likelihood and estimate filter coefficients α_{ik}. 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}, for single-mixture case, is given by

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

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} gives

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

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 aligned state to frame t of adaptation data) 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}. Once estimates of α_{ik} are obtained from adaptation data, they are used to transform all clean models to corrupted speech models. 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 0*th*-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.

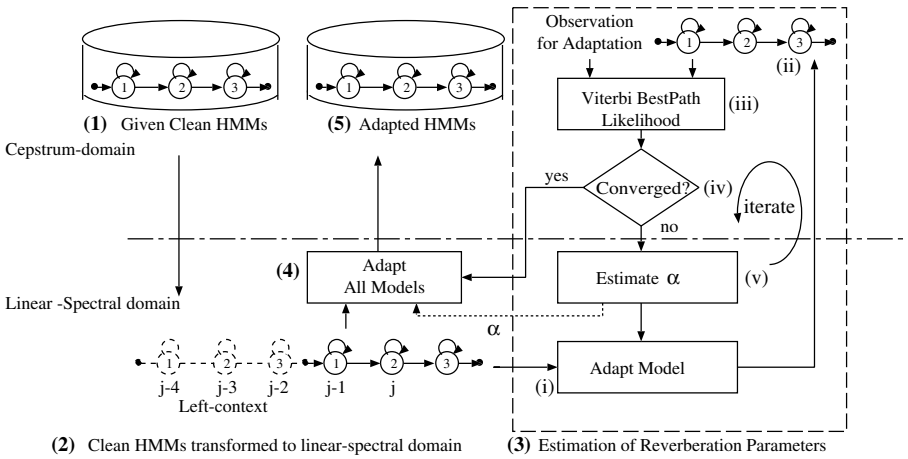

Fig. 2. Adaptation of model parameters for long convolutional distortion by maximum-likelihood state-filtering (MLSF)

Table 1. Experimental Results (Word Recognition Rate %)

Data	T_{60}	Clean	CMS	MLLR	MLSF(N=4)
Clean	–	97.9	97.6	97.9	97.9
E1B	310 ms	67.6	77.3	73.3	83.2
OFC	780 ms	44.8	47.5	49.0	72.5

For evaluation, convolutional distortion was simulated by convolving clean speech with impulse responses of the environment (viz. E1B and OFC) taken from RWCP Sound Scene Database in Real Environment. Recognition performance of distorted speech was evaluated with CMS and MLLR as well. For CMS, models were retrained with CMS performed training set data, and evaluation was done with CMS applied test set. For MLLR, global transformation matrix was estimated from adaptation data and used to adapt means. The result with MLLR varied with amount and phonetic coverage of adaptation data; the listed result in Table 1 is for 30 words of corrupted speech used as adaptation data. To evaluate the proposed maximum-likelihood based state filtering (MLSF) approach, only ten words of distorted speech was used as adaptation data to estimate α_{ik} with filter-order of $N = 4$, and states of left-context from biphones were considered for the adaptation.

Experimental results as listed under Table 1 show better performance of MLSF approach which demonstrates its effectiveness for convolutional distortion. The improvement has been obtained by using smaller amount of adaptation data than MLLR, and with longer reverberation time, the improvement with MLSF approach is more pronounced compared to other methods.

6. CONCLUSION

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

Future work includes evaluation of the method on large vocabulary continuous speech recognition (LVCSR) task with large Gaussian mixture models. Its capability to track and adapt to dynamic channel characteristics 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] S. Furui, "On the use of hierarchical spectral dynamics in speech recognition," in *Proc. ICASSP*, 1990, pp. 789–792.

[4] M. G. Rahim and B.-H. Juang, "Signal bias removal for robust telephone speech recognition in adverse environments," in *Proc. ICASSP*, 1994, vol. 1, pp. 445–448.

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

[6] 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.

[7] 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.

[8] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment independent speech recognition," in *Proc. ICASSP*, 1996, pp. 733–736.

[9] 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, pp. 171–185, 1995.

[10] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[11] 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.

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

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