Maximum Likelihood Based General Joint Adaptation to Noise and Long Reverberation *

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

1 Introduction

Background noise and reverberation present in speech signal degrade the performance of automatic speech recognition (ASR) systems to the extent of making them unusable for any applications. Several methods have been developed to deal with such distortion, ranging from various front-end methods like Cepstrum Mean Subtraction (CMS) [1], codeword-dependent cepstrum normalization (CDCN) [2] and RASTA [3] to different model based techniques like PMC [4], vector Taylor series based adaptation [5] and universal adaptation [6]. These methods have been reported to improve the performance of ASRs. However, most of them cannot perform reasonably well when reverberation time is much longer than analysis window-length, and in addition, background noise is also present. However, reverberation time longer than 100 ms is common in real-life environment like in office rooms, and most of the time, it is accompanied by background noise of one sort or another.

In this paper, we extend the work described in [7] to compensate model parameters *both* for additive noise and long reverberation present in speech signal, and investigate its capability to improve the performance of ASRs under noisy and reverberant environment. We introduce additive noise component term in compensation equation, and then estimate contributions from preceding states in maximum-likelihood manner from adaptation data, which approximates energy component contributed by preceding speech units due to long reverberation.

2 Effect of Additive Noise and Long Reverberation and Model Adaptation

The model for the environment with additive noise and reverberation is depicted in Fig 1. Considering the effect of long reverberation (reverberation time T_{60} longer than the analysis window-length) still convolutional (rather than multiplicative) in linear spectral domain, speech spectrum distorted by noise and long reverberation is approximated by

$$O(w_i,t) \approx H(w_i,t) * S(w_i,t) + N(w_i,t) \qquad (1)$$

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

*最尤基準に基づいた残響と雑音のためのモデル適応, 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)

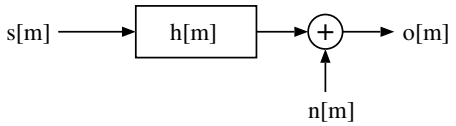

Figure 1: Model of environment with additive noise and long reverberation

Eq. 1 shows that the spectral parameters of distorted 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, due to long reverberation time. In [7], we accounted this by using mean vectors of preceding states in place of segments of preceding speech, and associating a parameter $\boldsymbol{\alpha}_i$ to each of these states to account for amount of their contributions to current state, which was estimated in maximum likelihood manner from few seconds of adaptation data.

The compensation to parameters essentially takes place in linear-spectral domain. To compensate for additive noise, we introduce mean and covariance of noise into the compensation equation. Noise mean and covariance are estimated from noise samples obtained during non-speech activity. After transforming the cepstrum-domain model parameters to linear spectral domain, the model parameters of state j, viz. mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, for distorted speech are estimated as:

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

where k, l represent dimensions of parameters, α_{ik}s are filter coefficients, and subscript lin signifies linear domain parameters. From preceding states, only composite means (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. Covariance matrix can be retained unchanged, or can adapted as in Eq. 3 (gain multipliers to account level difference can be also introduced in the equation).

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

3 Maximum-Likelihood Estimation of State Filter coefficients

Model λ_O for distorted speech is composed by using clean speech model λ_S, noise model λ_N and estimated parameters $\boldsymbol{A} = \{\boldsymbol{\alpha}_0, \ldots, \boldsymbol{\alpha}_{N-1}\}$. The parameters α_{ik} are estimated by maximizing Viterbi-likelihood score $P(\boldsymbol{O}, \boldsymbol{q}|\boldsymbol{A}, \lambda_S, \lambda_N)$ or $P(\boldsymbol{O}, \boldsymbol{q}|\lambda_O)$ of adaptation 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

$$\hat{\alpha}_{ik} = \arg\max_{\alpha_{ik}} P(\boldsymbol{O}|\boldsymbol{\alpha}_{0,\ldots,N-1}, \lambda_S, \lambda_N). \quad (4)$$

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

$$\alpha_{ik}(p) = \alpha_{ik}(p-1) + \epsilon \frac{\partial \log\left(P(\boldsymbol{O}|\boldsymbol{\alpha}_{0,\ldots,N-1}, \lambda_S, \lambda_N)\right)}{\partial \alpha_{ik}} \quad (5)$$

where ϵ is scaling factor. The details of such a maximization is given in [7].

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 with T_{60} = 310 ms and OFC with T_{60} = 780 ms) taken from RWCP Sound Scene Database in Real Environment. Fan noise from JEIDA database was added to reverberant speech signal at SNR of 20 dB.

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. A single state noise HMM was trained from noise sample and compensation to mean only was considered during estimation of α_{ik} and model adaptation.

The experimental result without fan noise case is listed in Table 1, whereas that with fan noise is listed in Table 2, which shows degradation in word accuracy with addition of fan noise with clean model. Though CMS does improve word accuracy for noisy and reverberant speech, the improvement with joint compensation (JC) by proposed maximum-likelihood approach is significantly better. Further, with longer reverberation time, the method gives significantly higher improvement compared to CMS in both cases of with and without fan noise.

Future work includes evaluation of the method on large vocabulary continuous speech recognition task and with different types of noise at varying SNRs.

Table 1: Expt. Result (Word Recognition Rate %) for environment with long reverberation for clean model, CMS and Maximum Likelihood State Filtering (MLSF) [7] approach

Data	Clean	CMS	MLSF
Clean	97.9	97.6	97.9
E1B (310 ms)	67.6	77.3	83.2
OFC (780 ms)	44.8	47.5	72.5

Table 2: Expt. Result (Word Recognition Rate %) for environement with *both* additive noise and long reverberation

Data	Clean	CMS	JC
Clean	97.9	97.6	97.9
E1B (310 ms) + 20 dB fan	15.1	44.7	87.3
OFC (780 ms) + 20 dB fan	12.8	18.85	67.1

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

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

[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] 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.

[6] 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.

[7] C. K. Raut, T. Nishimoto and S. Sagayama, "Model adaptation for long convolutional distortion by maximum likelihood based state filtering approach," in *Proc ICASSP*, 2006, to appear.