INSTANTANEOUS ENVIRONMENT ADAPTATION TECHNIQUES BASED ON FAST PMC AND MAP-CMS METHODS

Tetsuo KOSAKA Hiroki YAMAMOTO Masayuki YAMADA and Yasuhiro KOMORI

Media Technology Lab., Canon Inc.
53, Imaikami-cho, Nakahara-ku, Kawasaki-shi, Kanagawa 211, Japan

ABSTRACT

This paper proposes instantaneous environment adaptation techniques for both additive noise and channel distortion based on the fast PMC (FPMC) and the MAP-CMS methods. The instantaneous adaptation techniques enable a recognizer to improve recognition on a single sentence that is used for the adaptation in real-time. The key innovations enabling the system to achieve the instantaneous adaptation are: 1) a cepstral mean subtraction method based on maximum a posteriori estimation (MAP-CMS), 2) real-time implementation of the fast PMC [5] that we proposed previously, 3) utilization of multi-pass search, and 4) a new combination method of MAP-CMS and FPMC to solve the problem of both channel distortion and additive noise. Experiment results showed that the proposed methods enabled the system to perform recognition and adaptation simultaneously nearly in real-time and obtained good improvements in performance.

1. INTRODUCTION

To realize practical speech recognition systems, high accurate systems in a wide variety of noise environments are required. On a telephone speech recognition, problems are both additive noise like background noise and channel distortion caused by the difference in telephone line characteristics. Especially for the telephone speech recognition, the noise environment greatly differs according to circumstances. In order to solve this problem, we propose instantaneous environment adaptation techniques for both additive noise and channel distortion. Instantaneous adaptation enables a recognizer to improve recognition on a single sentence that is also used for the adaptation. The instantaneous adaptation must have the following three characteristics:

1) Unsupervised adaptation is possible.
2) Improvement in performance must be attained with short-time calibration data (e.g. single utterance).
3) Both recognition and adaptation are carried out simultaneously nearly in real-time.

It is well known that Cepstral Mean Subtraction (CMS) [2] is one of the accurate methods of channel normalization. Parallel Model Combination (PMC) [3] has been proposed for additive noise. Both methods nearly satisfy the conditions of 1) and 2). However, both methods are not suitable for real-time implementation. To solve the problem, we propose a MAP-CMS algorithm and real-time implementation of a fast PMC (FPMC). Furthermore, a new combination method of MAP-CMS and FPMC is proposed.

The conventional CMS method is not suitable for instantaneous adaptation because CMS can not be synchronized with the recognition procedure. Rahim et al. proposed a sequential CMS and sequential SBR methods to facilitate real-time implementation[6]. However, there is a tendency that useful speech information is removed during the initial part of the utterance with these methods. We propose MAP based CMS (MAP-CMS) in which the MAP estimation complements the lack of training data during an initial part of the utterance. We also propose some implementation of MAP-CMS.

Previously, we proposed a Fast PMC (FPMC) algorithm [5] in which computational cost was saved with almost no degradation of recognition performance. This method quite differs to data-driven PMC (DPMC) [4] in which a way to reduce the PMC computation amount was introduced. In order to realize the instantaneous adaptation by using FPMC, some implementation of FPMC based on a tree-trellis based search [7] are proposed.

Furthermore, the new combination method of MAP-CMS and FPMC is proposed to overcome adverse conditions which include both channel distortion and additive noise.

2. MAP-CMS FOR INSTANTANEOUS ADAPTATION

In this section, we propose a MAP (Maximum a Posteriori Estimation) based CMS method (MAP-CMS) to realize frame-synchronous CMS. CMS on input parameter sequence represented in cepstral domain is calculated as

$$\hat{x}_n = x_n + \mu_d - \mu \tag{1}$$

where x_n is an observation vector at the $n-th$ frame, μ_d is the mean of training sample, μ is the mean of observation and $\hat{x}_n$ is the normalized vector. In the case of frame-synchronous CMS, μ cannot be estimated accurately because of the lack of training sample just after a starting of utterance. We employ MAP estimation to improve estimation accuracy of μ. MAP estimation uses information from an initial model as a priori knowledge to complement the lack of training data. Assume the prior pdf is Gaussian with mean μ_o and variance σ_o^2. The MAP estimates of the mean μ_{MAP} are given by[1]

$$\mu_{MAP} = \frac{n}{n+\tau} m + \frac{\tau}{n+\tau} \mu_o \tag{2}$$

where m is the sample mean ($m = (1/n)\sum_{k=1}^{n} X_k$) and also the Maximum Likelihood estimate, n is the number of training samples observed for the corresponding Gaussan, and τ indicates a relative balance between the prior and training data. Here we employ Gaussian distribution estimated from training data $N(\mu_d, \sigma_d^2)$ as the prior. Substituting Eq.(2) for Eq.(1), $\hat{x}_n$ is given by

$$\begin{aligned}\hat{x}_n &= x_n + \mu_d - \mu_{MAP} \\ &= x_n + \mu_d - (\frac{n}{n+\tau}m + \frac{\tau}{n+\tau}\mu_d) \\ &= x_n + \frac{n}{n+\tau}(\mu_d - \frac{1}{n}\sum_{k=1}^{n} x_k) \end{aligned} \quad (3)$$

If the number of observation samples is very small (i.e. $n \simeq 0$), then almost no transformation is carried out, and when n is infinite, this equation is equal to equation of conventional CMS (i.e. Eq. (1)).

We propose three types of implementation of MAP-CMS as follows:

forward MAP-CMS In a forward search of a tree-trellis based decoder [7], MAP-CMS is carried out frame-synchronously, and no output probability is recalculated in backward.

backward MAP-CMS In the forward search, cepstral mean is calculated but MAP-CMS is not carried out. In the backward search, subtraction is carried out by using Eq. (3). In this method, the cepstral mean can be estimated accurately because it is estimated from the whole utterance.

forward-backward MAP-CMS MAP-CMS is carried out in both forward and backward. The backward MAP-CMS is expected to complement the estimation of cepstral mean which may not be accurate in the initial part of the input utterance in the forward search.

In the cases of the backward MAP-CMS and the forward-backward MAP-CMS, output probabilities are different between forward and backward search. Therefore, the A^* condition for the optimality is no longer satisfied. However, keeping enough N-best stack size is considered to save the search error. It is just the same with a backward FPMC which is described in Section 3.2.

3. REAL-TIME IMPLEMENTATION OF FAST PMC

In this section, we describe our recent work on real-time implementation of a fast PMC (Parallel Model Combination) algorithm which we previously proposed[5].

3.1. Fast PMC

The basic PMC algorithm [3] generates the cepstrum-based noise corrupted HMM from the noise HMM and the speech HMM, each of which is separately modeled. In order to realize a fast PMC (FPMC) noise adaptation, we make the following assumptions: 1) The noise corrupted position of each distribution can be determined from the difference between the close distributions and the composite distribution before PMC by taking account of the area corruption. 2) The noise corrupted area of each distribution can be determined from the area ratio of the composite distribution before and after PMC.

The image of the proposed method is shown in Fig. 1. In the basic PMC, all distributions must perform the PMC-processing, while the proposed method requires a single PMC-processing per a composite distribution. The algorithm of the FPMC is shown as follows:

1. Group close distributions (μ_m, σ_m^2) and create a composite distribution (μ_c, σ_c^2) per group:
$$\mu_c = \sum_{m \in G} w_m \mu_m \quad (4)$$
$$\sigma_c^2 = \sum_{m \in G} w_m \sigma_m^2 + \sum_{m \in G} w_m (\mu_m - \mu_c)^2 \quad (5)$$
where w indicates weights and G indicates groups. In this paper, the group is the state of HMM.
2. Calculate vectors of the difference between distributions (μ_m, σ_m^2) in the group and the composite distribution (μ_c, σ_c^2) of the group.
3. Perform PMC-processing on the composite distribution [3].
4. Calculate the noise corrupted position and an area of each distribution by the difference between each distribution and the composite distribution before PMC, and the area ratio of the composite distribution before and after PMC, using the next equations:
$$\hat{\mu}_{m,S+N} = \hat{\mu}_{c,S+N} + (\mu_{m,S} - \mu_{c,S})\ (\hat{\sigma}_{c,S+N}/\sigma_{c,S}) \quad (6)$$
$$\hat{\sigma}_{m,S+N} = \sigma_{m,S}\ (\hat{\sigma}_{c,S+N}/\sigma_{c,S}) \quad (7)$$
while $_{S+N}$ indicates noisy speech and $_S$ indicates clean speech and μ, σ before adaptation and $\hat{\mu}, \hat{\sigma}$ after adaptation.

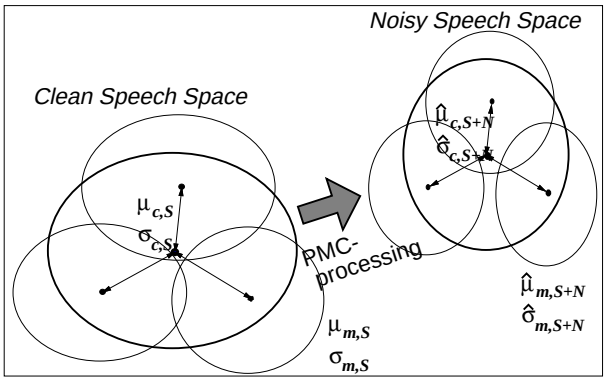

Figure 1: Image of Fast PMC Processing

3.2. Instantaneous Adaptation Using FPMC

A computational cost can be saved drastically by using FPMC algorithm. It can save around 2/3 of basic PMC computation amount with almost no degradation of recognition performance when right context models are used [5].

However, it takes several seconds to adapt models even by using FPMC algorithm. To realize real-time recognition, we propose following two types of implementation:

Forward FPMC In a forward search of the tree-trellis based decoder [7], not all acoustic models but selected models which are required in the linguistic search are adapted by FPMC algorithm. In a backward search, output probability is not recalculated. Using this method, redundant computations can be avoided without degradation of recognition performance, and a backward search is very fast because of no output probability calculation.

Backward FPMC In a forward search, acoustic models are not adapted. In a backward search, models are adapted by FPMC algorithm, and output probabilities are recalculated by using adapted models. Because adaptation is carried out only in the backward search, the recalculation cost of output probabilities is small.

4. COMBINATION OF MAP-CMS AND FPMC

In this section, a new combination method of MAP-CMS and FPMC is proposed to overcome adverse conditions which include both additive noise and channel distortion.

It is difficult to do both MAP-CMS and FPMC simultaneously in the forward search. Since a subtraction amount varies frame by frame with MAP-CMS processing, the FPMC processing must be done at every frame. In our proposed method, MAP-CMS is carried out in the forward search and both MAP-CMS and FPMC are carried out in the backward search. Since only channel distortion is removed in the forward search with this method, recognition performance may drop in low SNR. Then a spectral subtraction method (SS) is adopted as a pre-processing. The algorithm of the combination method is as follows:

1. Carry out SS before parameter calculation.
2. Normalize input parameters by using MAP-CMS in the forward search.
3. Estimate parameters of noise HMM which consists of single Gaussian pdf. Note that the parameters of HMM must be normalized as follows:

$$\hat{\mu}_p = \mu_p + \mu_d - \mu_{MAP} \tag{8}$$

   where μ_p is the mean of noise HMM, and $\hat{\mu}_p$ is the normalized mean.
4. Carry out MAP-CMS and FPMC in the backward search.

5. RECOGNITION EXPERIMENTS

5.1. Experimental Conditions

The proposed methods were evaluated in Japanese sentence recognition. Conditions are briefly shown in the Table 1. The tasks were 1,000 vocabulary size continuous speech recognition uttered by 10 speakers. Acoustic models used here were right context phone HMMs of 3-state 6-mixture. The total number of HMMs is 262. Stack depth for the tree-trellis based search was 35. Noise data of 1.0 second was used for PMC adaptation. In the experiments of MAP-CMS, channel distortion was artificially added by using BPF (300 - 3,200Hz). In the experiments of FPMC, computer room noise was artificially added to evaluation data. Both channel distortion and additive noise were added in the same way for the evaluation of the combined method. The WS used here was HP/K260EG (SPECfp95 = 19.4).

Table 1: Experimental Conditions

Acoustic Analysis	sampling rate: 8kHz, frame period: 10ms, hamming window: 25.6ms LPC-Mel-Cep(12 dimension) + ΔCep(12 dimension) + Δpower
Training Data	ASJ+ATR speech data 104 speakers, 20840 utterances
Evaluation Data1	CANON speech database 1,004 words, perplexity 30.2 10 speakers, 500 sentences

5.2. Results of MAP-CMS

Proposed three methods were compared with two types of baseline methods. One is no adaptation as a lower limit baseline experiment, the other is CMS using the whole of each utterance as adaptation data for an upper limit baseline. The recognition results for various τ values are shown in Fig. 2. The results of the comparison in recognition time are shown in Fig 3. In every case of MAP-CMS, calculation time for the forward search is much smaller than average duration of input utterances ($= 2.85sec$). This means real-time computation can be done in the forward search. Difference in recognition rate between the forward-backward MAP-CMS and upper limit baseline was 0.6 % at $\tau = 20.0$. In this case, backward calculation time was 0.26 sec. Compared with the conventional CMS, 3/4 of the computation was saved. In comparison among three methods, the forward-backward method showed the best recognition rate. The recognition performance of the backward method is not good because correct candidate tends to be pruned in the forward search. Note that the forward method shows the best performance in recognition time because no output probability is recalculated in the backward search.

5.3. Results of FPMC

Proposed two methods were compared with two types of baseline methods. One is no adaptation as a lower limit, the other is conventional PMC as an upper limit. The recognition results are shown in Fig. 4. In comparison between PMC and FPMC, the difference in recognition performance was very small, even though FPMC was approximation of PMC. The recognition time was 2.84sec (forward:2.79 sec + backward:0.05 sec) by using the forward FPMC at SNR = 20dB. Then recognition can be carried out almost in real-time by using the forward FPMC because the average duration of input utterance is 2.85 sec. Since it takes 5.94 sec for only adaptation by using the conventional PMC, it is not suitable for real-time recognition. When the backward FPMC was employed, time of the backward search was long (0.97sec at SNR20dB) and recognition rate of the backward FPMC was worse than that of the forward FPMC.

5.4. Results of the Combination Method

The following four types of methods were compared.

1) **NONE** No adaptation.

2) **forward FPMC** FPMC in the forward search.

3) **SS+for-back MAP-CMS** Forward-backward MAP-CMS was carried out with spectral subtracted parameters.

4) **SS+for-back MAP-CMS+back FPMC** Backward FPMC was added to the above method.

The results are shown in Table 2. In the method 2), the recognition rate was not good because only additive noise were considered. Methods 3) and 4) indicated good results because both additive noise and channel distortion were considered in these methods. The performance of 4) was better than that of 3). This means the additional backward FPMC is effective in spite of using spectral subtracted parameters.

6. CONCLUSION

This paper proposed instantaneous environment adaptation techniques for both channel distortion and additive noise based on MAP-CMS and FPMC. The forward-backward MAP-CMS could save 3/4 of the recognition time with almost no degradation of recognition performance. The forward FPMC also saved the recognition time and the recognition can be carried out almost in real-time. Furthermore the combination method of MAP-CMS and FPMC was proposed. As results of the evaluation experiments, the effect of the combination method was proved.

7. REFERENCES

[1] Duda R.O. and Hart P.E.: Pattern Classification and Scene Analysis. New York: Wiley, 1973.

[2] Furui S.: Cepstral analysis technique for automatic speaker verification, IEEE ASSP, 29, pp. 254-272 (1981.4).

[3] Gales M.J., et al.: An improved approach to the hidden Markov model decomposition of speech and noise, ICASSP92, pp.233-236, 1992.

[4] Gales M.J., et al.: A fast and flexible implementation of Parallel Model Combination, ICASSP95, pp.I-133-136,1995-5.

[5] Komori Y., Kosaka T., Yamamoto H., Yamada M.: Fast Parallel Model Combination Noise Adaptation Processing, Proc. of Eurospeech97, pp. 1527-1530 (1997.09).

[6] Rahim M.G. and Juang B.-H.: Signal Bias Removal by Maximum Likelihood Estimation for Robust Telephone Speech Recognition, IEEE Trans. on Speech, Audio Processing, Vol. 4. No. 1, pp.19-30 (1996.1).

[7] Soong F. et al.: Tree-trellis based fast search for finding the N best sentence hypotheses in continuous speech recognition, ICASSP'91, pp.705-708, 1991.

Table 2: Recognition Results of Combination Methods (%)

	Methods/SNR	10	15	20	30
1)	NONE	0.6	8.4	29.4	58.8
2)	forward FPMC	8.6	18.4	34.6	58.4
3)	SS+for-back MAP-CMS	16.6	37.8	56.2	71.8
4)	SS+for-back MAP-CMS +back FPMC	21.4	41.2	60.8	72.6

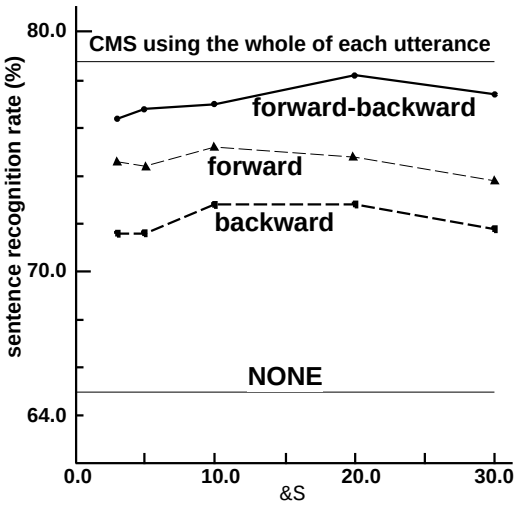

Figure 2: Recognition Results of MAP-CMS

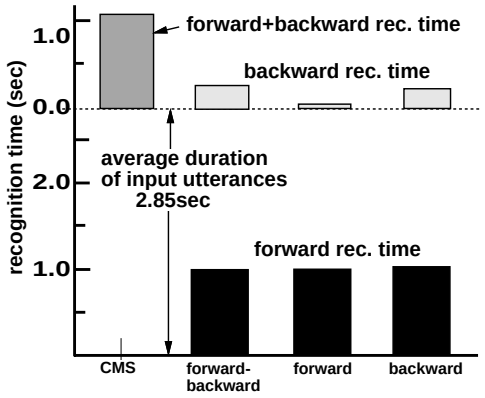

Figure 3: Recognition Time of MAP-CMS($\tau = 20.0$)

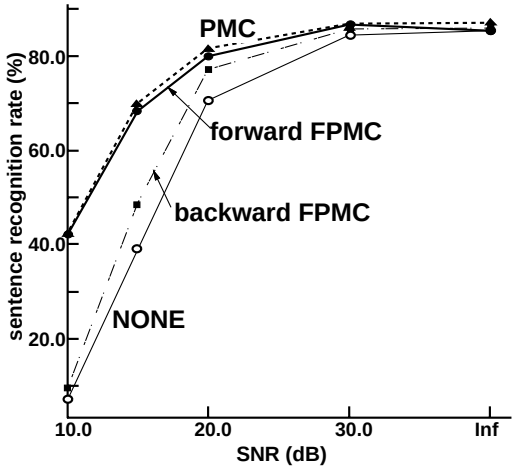

Figure 4: Recognition Results of FPMC