Speech Recognition in Noise Using On-line HMM Adaptation

TungHui Chiang

Advanced Technology Center (ATC)
Computer & Communication Laboratories (CCL)
Industrial Technology Research Institute (ITRI)
Chutung, Hsinchu, Taiwan 310, R.O.C

ABSTRACT

In this paper, a novel two-stage framework is proposed to copy with speech recognition in adverse environment. First, an on-line HMM composition method which compensates HMMs making use of the on-line testing utterances is proposed in the first stage. By using the proposed method, the dynamic change of environmental noise in each utterance can be well handled. In addition, a classifier trained by using a discriminative learning procedure is incorporated in the second stage to enhance system's discrimination capability. Since the recognition and adaptation processes are carried out in the same session in an unsupervised fashion, this proposed two-stage framework is suitable for practical uses.

1. INTRODUCTION

The parallel model combination (PMC) method [1, 2] has been shown to be effective for speech recognition in noise. Especially, the PMC method, unlike other environmental adaptation methods [3], needs no speech data in the testing environment for parameter compensation. By using the PMC method, however, the environment noise should be collected in advance to construct a noise HMM. Afterwards, the original HMMs are combined with the noise HMM to build the environment-dependent HMMs.

In general, the PMC method performs well as the adaptation environment is the same as the testing environment. However, the environmental noises usually change from time to time in real applications. If the environment changes slowly, the PMC compensating procedure can be repeated at a period of time. Nevertheless, if the environment changes rapidly or the adaptation environment differs far from the testing environment, the PMC method usually fails to achieve high performance due to serious model mismatch.

An illustration of model mismatch in training and testing under different noise conditions is shown in figure 1. Compared to the model space in a mismatched condition, e.g., S2, the compensated model space S1 in the matched condition is considered to be "closer" to the desired model space S3.

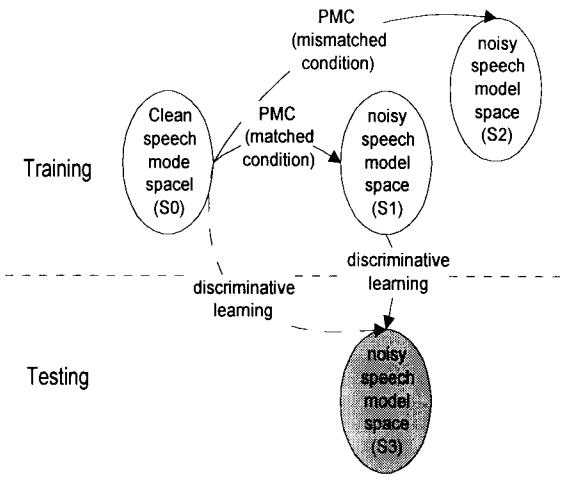

Figure 1. Model space mismatch in training and testing under different noise conditions

In addition, the PMC compensation processes [1, 2] obtain the model parameters by using either Gaussian integration or maximum likelihood estimation (MLE). Generally, the models obtained in thus ways do not necessarily guarantee to minimize the error rate of the test data because neither discrimination nor robustness issues are considered in the entire compensation process.

Conceptually, to minimize the test set error, the model space should be adapted as close as possible to the model space of the testing data, i.e., S3 in figure 1. For this purpose, two model adaptation methods can be performed. One is to adapt model parameters directly from the model space trained by clean speech. Usually, a set of contaminated adaptation data are needed for this kind of adaptation method to achieve a satisfactory result. However, the requirement of contaminated adaptation data makes this method inadequate to most applications.

The other way for model adaptation is to start the learning process from the PMC compensated models. An advantage for adopting this method lies in the fact that the model space compensated by the PMC method is much closer to that of testing data, compared to that trained by clean speech. To start an adaptation procedure from a better initial point can, in general, prevent from being trapped to a poor local optimal point. Therefore,

performing adaptation from the PMC compensated models is more likely to attain a better result.

Motivated by the above-mentioned concerns, a two-stage framework which integrates on-line model composition and unsupervised model adaptation in a same session is proposed. The block diagram of the proposed framework is illustrated in figure 2. First, the recognized testing utterances are used for noise model re-estimation. The re-estimated noise model is then used for model composition to better track the environmental change. Afterwards, following the first stage, a discriminative classifier is adopted as the second stage to enhance the discrimination capability of the overall system. To minimize the error rate, this classifier is trained by using MCE-based adaptive learning procedures [4,5,6]. Experiments have shown the proposed two-stage approach is quite affective and robust against dynamic change of environment noise.

2. THE ON-LINE PARALLEL MODEL COMPOSITION

To better track the dynamic change of testing environment, the on-line parallel model composition method first uses the testing utterances themselves for noise model reestimation. Afterwards, the estimated noise model is combined with the clean speech models using the well known PMC procedure. Implementation of this method consists of the following steps.

Step 1. Preliminary Recognition and Alignment:

To recognize the input utterances by using the viterbi decoding scheme; then backtrack the state sequences to obtain the frames which are aligned to silence (noise) states.

Step 2. Noise Model Re-estimation:

To estimate the noise model by using the information provided by the frames determined in the step 1. Smoothing methods, such as the deleted interpolation method, could be adopted in this step. In this paper, the recursive ML estimation is adopted.

Let $\Lambda(n)$ stand for the parameter estimated from the noise portions of the previous utterances, which contain n number of frames, and $\Lambda(k)$ for the parameter estimated from the noise portions of the current utterance containing k frames. The re-estimated noise model, denoted by $\Lambda(n+k)$, can be repressed as an interpolation of $\Lambda(n)$ and $\Lambda(k)$. For instance, by assuming single Gaussian mixture for each state in the noise model, the mean value of the i-th component of a particular state can be represented as follows:

$$\mu_i(n+k) = \frac{n}{n+k}\mu_i(n) + \frac{k}{n+k}\mu_i(k).$$

Meanwhile, the corresponding variance can be re-estimated according to the following expression:

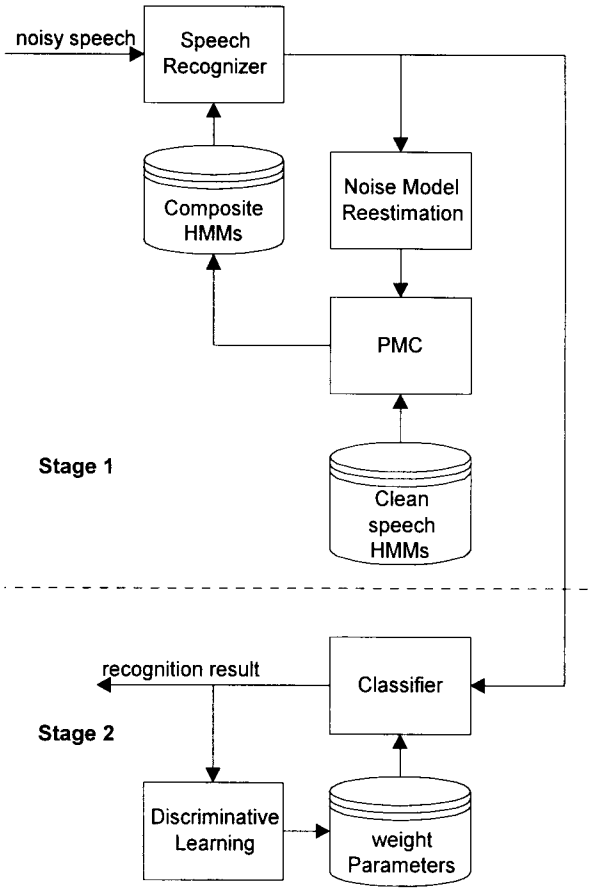

Figure 2. System Block Diagram

$$\sigma_i(n+k) = \left[\frac{n}{n+k}m_i^2(n) + \frac{k}{n+k}m_i^2(k)\right] - \left[\mu_i(n+k)\right]^2,$$

where $m_i^2(n) = \frac{1}{n}\sum_{j=1}^{n} x_i^2(j)$ denotes the estimated second moment from n samples.

Step 3. Model Composition:

To compose the original HMMs with the re-estimated noise model using the PMC method.

It is noticed that no matter what the recognition results conduct, only the aligned noise portions of the input utterances are of interest in the step 2. Therefore, this framework is able to perform under an unsupervised condition. In addition, having the unsupervised learning capability, the proposed on-line PMC method, unlike the conventional speech adaptation/recognition scheme, can carry out the model adaptation and recognition processes in the same session. Thus, the on-line testing utterances can be used for adaptation to cope with the dynamic change of environment in each testing utterance.

3. DISCRIMINATIVE LEARNING

Regarding to the PMC approaches [1, 2], the

models are compensated by using either Gaussian integration or maximum likelihood estimation (MLE). The models obtained in thus ways do not necessarily guarantee to minimize the error rate of the test data because neither discrimination nor robustness issues are considered in the entire compensation process.

To minimize the error rate, a classifier is adopted in the second stage. In this classifier, the discrimination function $g_j(O)$ of the input utterance O with respect to the j-th word W_j is defined in terms of weighted HMM [4] as follows:

$$g_j(O) = \sum_{k=1}^{N_j} w_{j,k} \cdot SC_{j,k},$$

where N_j stands for the total number of distinct states representing W_j; $SC_{j,k}$ for the accumulated log probabilities assigned to the k-th state; $w_{j,k}$ for the corresponding state weight. Initially, $w_{j,k}$ for all j and k are set to one. The word W_{opt} is considered as the recognition result if this word leads to the maximum value of the discrimination function; that is:

$$W_{opt} = \arg\max_{W_j} g_j(O).$$

Furthermore, based on the discrimination function, a distance function measuring the degree of miss-recognition between two competing candidates W_q, W_r is defined as follows:

$$d_{qr}(O) = g_q(O) - g_r(O).$$

In a supervised learning procedure, W_q stands for the correct candidate, and W_r for the top competitor. Hence, a recognition error occurs in case $d_{qr}(O) < 0$. On the other hand, when the learning procedure is performed in an unsupervised fashion, W_q stands for the top candidate and W_r for the top second candidate.

Afterwards, a loss function as an approximation of error function needs to be defined. Currently, the loss function is defined as follows:

$$\ell(d) = \begin{cases} \tan^{-1}\left(\frac{d}{d_0}\right) & d < 0, \\ 0 & otherwise, \end{cases}$$

where d_0 is a small positive constant. To taking robustness issue into consideration while the probabilistic decent theory is applied, the adjustment of parameters $\Delta\Lambda_t$ at time t would satisfy the following equations [4, 6]:

$$\Lambda_{t+1} = \Lambda_t + \Delta\Lambda_t, \quad if\ d(O) < \tau,$$
$$\Delta\Lambda_t = -\varepsilon(t) U \nabla \bar{R},$$

where $\tau(\tau > 0)$ is a preset margin; $\varepsilon(t)$ is a decreasing function of t; U is a positive-definite matrix, which is assumed to be an identity matrix in this paper; $\bar{R}$ is the averaged loss function. Readers who are interested in the learning algorithm are pleased to refer to [4, 5, 6] for details.

4. EXPERIMENTAL EVALUATION

4.1 Experimental Setup

In following experiments, the acoustic feature vectors are extracted from the 8KHz sampled data every 10msec. Each feature vector is composed of 12-order mel-scaled cepstral coefficients and the corresponding delta cepstral coefficients. A total of 58 phone-like units (PLU) HMMs are used as acoustic units for speech recognition. Each PLU is modeled by a 3-state 4-mixture continuous density HMM. The training of clean speech PLU HMMs was carried out by using the following two speech databases:

[1] utterances from 90 speakers (50 male and 40 female), each speaking 408 Chinese 408 base syllables.

[2] utterances from 16 speakers (5 male ad 11 female), each speaking 479 poly-syllable words.

In addition, a task of recognizing 120 person names is used for evaluation. In the following experiments, the testing set consists of 600 utterances recorded from 5 male speakers. These utterances are mixed with the NOISEX-92 *speech noise* data at 4 different levels, ranging from 0dB to 18dB.

First, to investigate the effect of noise type mismatch on model compensation, the experiments are performed under the following two conditions:

(1) *noise-type matched condition*, i.e., the noise type in the training environment is the same as that of the testing environment.

(2) *noise-type mismatched conditions*, i.e., the noise types in the training environment are different from that of the testing environment. Here, three types of NOISEX-92 noise source, including *car noise, machine gun noise and lynx noise* are used for model composition. Those composite models are then used to recognize the test data contaminated by *speech* noise.

4.2 Results and Discussions

Table 1 summarizes the performances at different noise levels using the clean speech models, PMC compensated models in the matched and mismatched conditions, and the on-line PMC. As shown in table 1, we find that the PMC compensated models, even with very mismatched noise, e.g., the machine gun noise, achieve better performances in comparison with the model trained by clean speech. It is also found in table 1 that compensation with the appropriate noise would achieve a better result. For instance, the PMC method in the matched condition attains the performance 3-17% better than those obtained in the mismatched conditions when SNR is equal to 6dB. Particularly, as regard to the proposed on-line PMC method, it achieves the performance almost as high as that obtained by the PMC method in the matched condition. The promising result

demonstrates that the proposed on-line PMC method is quite robust against environment changes.

	0dB	6dB	12dB	18dB
Clean	12.5	35.8	71.6	84.2
PMC (Car)	17.5	54.2	79.2	89.2
PMC (Gun)	17.5	40	71.7	87.5
PMC (Lynx)	17.2	51.7	82.5	91.7
PMC (Speech)	18.3	57.5	82.5	91.7
On-line PMC	21.7	55.8	83.3	90.8

Table 1. Performances of various PMC-based methods.

Furthermore, to evaluate the proposed two-stage approach, the learning procedures in unsupervised and supervised manners are applied to adjust the weight parameters of the classifier in the second stage. It is noticed that users are supposed to provide the system the correct answer for each utterance in the supervised learning procedure. However, in the unsupervised learning mode, the system assumes that the recognition result is the correct candidate.

The results of the two-stage approach are listed in table 2, where UDL and SDL correspond to the unsupervised and the supervised discriminative learning, respectively. Note that the result with the supervised learning servers the upper bound which the proposed two-stage approach can achieve.

	0dB	6dB	12dB	18dB
On-line PMC+UDL	21.7	58.3	83.5	91.7
On-line PMC+SDL	22.5	71.7	90.8	95
PMC(speech)+UDL	21.7	58.3	83.5	92.5
PMC(speech)+SDL	22.5	71.7	90.8	95

Table 2. Performances of the two-stage framework.

Compared to the result in table 1, the unsupervised learning provides limit improvement. The reason why the unsupervised learning fails to get much improvement lies in the fact that the learning procedure tends to be trapped to a poor local optimum easily without a good supervision. This problem is particularly serious when SNR is low.

In contrast, if the discriminative learning is performed in a supervised manner, the performance is improved significantly at wide range of noise levels. Therefore, appropriate incorporation of confidence measures into the learning procedure might provide a way to bridge the broad gap between the unsupervised and the supervised learning procedures.

Moreover, for comparison's sake the performances of applying the discriminative learning procedures starting from the PMC compensated models are also listed in table 2. The corresponding results show almost the same as those with the proposed two-stage approach. Again, those results demonstrate the superiority of the proposed approach to cope with the problems of speech recognition in noise.

5. SUMMARY

In this paper, a two-stage framework is proposed to copy with speech recognition in adverse environment. In the first stage, an on-line PMC method which compensates HMMs making use of the on-line testing utterances is proposed. By using the proposed method, the dynamic change of environmental noise in each utterance can be well handled. In addition, a linear classifier trained by using a discriminative learning procedure is incorporated in the second stage to enhance system's discrimination capability. Since the recognition and adaptation processes are carried out in the same session in an unsupervised fashion, this proposed two-stage framework is suitable for practical uses. Even thought the unsupervised discriminative learning shows limit improvement, the supervised learning procedure provides a significant improvement. Therefore, to find out a better way for applying the unsupervised learning procedure, such as providing a confidence measure, will be our future research.

REFERENCES

[1] M. Gales and S. Young, "Cepstral parameter compensation for HMM recognition in noise," Speech Communication, pp. 231-239, Vol 12, 1993.

[2] M. Gales and S. Young, "A fast and flexible implementation of parallel model combination," pp. 133-136, Proceedings of ICASSP'95, pp. 231-239, Vol 12, 1993.

[3] G. Zavaliagkos, R. Schwartz, and J. Makhoul, "Batch, incremental and instantaneous adaptation techniques for speech recognition," pp. 676-679, Proceedings of ICASSP'95, 1995.

[4] K. Y. Su and C. H. Lee, "Speech recognition using weighted HMM and subspace projection approaches," IEEE Trans. On speech and Audio Processing, vol. 2, no. 1, pp. 69-79, Jan. 1994.

[5] B. H. Juang and S. Katagiri, "Discriminative learning for minimum error classification, " IEEE Trans. On Acoustic, Speech, and Signal Processing, vol. 40, no. 12, pp. 3043-3054, Dec. 1992.

[6] T. H. Chiang, Y. C. Lin, K. Y. Su, "On jointly learning the parameters in a character-synchronous integrated speech and language model, " IEEE Trans. On speech and Audio Processing, vol. 4, no. 3, May 1996.