A DIVERGENCE-BASED MODEL SEPARATION

Chao-Shih HUANG
Philips Research East Asia, Taipei
Department of Electrical Engineering, National Tsing Hua University, Hsinchu
joseph.huang@philips.com

Hsiao-Chuan WANG
Department of Electrical Engineering, National Tsing Hua University, Hsinchu
hcwang@ee.nthu.edu.tw

ABSTRACT

In this paper, a divergence-based training algorithm is proposed for model separation, where the relative divergence between models is derived from Kullback-Leibler (KL) information. We attempt to improve the discriminative power of existing model while the environment-matched training data is not available. It could be applied to improve the model discrimination after model-based compensation technique is performed for robust speech recognition. Traditionally, the model training is based on data driven such as maximum likelihood (ML) estimation or discriminative training. Compared to ML training, the minimum classification error (MCE) objective in discriminative training leads significant gain in accuracy. We attempt to improve model discrimination based on an approximate classification error analysis, relative divergence. We found that the smaller the relative divergence is, the more discriminative powers of the two models are. In the proposed algorithm, we try to directly obtain the discriminant function for model training from the relative divergence. Thus, the model parameters can be adjusted based on minimum relative divergence. Experimental results demonstrate that the divergence-based model separation method can achieve better recognition performance.

1. INTRODUCTION

The hidden Markov model (HMM) paradigm is widely used in automatic speech recognition systems [1]. The speech model of HMM-based recognizer is traditionally constructed in the sense of maximum likelihood (ML) estimation. However, the ultimate objective is not directly to model the distributions of speech but rather to minimize the recognition error rate. Note that the ML criteria does not minimize the recognition error rate. The other attempt is so called discriminative training [2,3] by directly targeting the minimum classification error (MCE) objective. The discriminant function in MCE algorithm is estimated from the classification of input data via speech models. Compared to ML-based recognizer, the discriminative training has been proved the significant gains in accuracy and robustness. Both ML and MCE training schemes are data driven.

In this paper, our motivation is to improve the model discrimination without training data. For example, how can we improve the model discrimination for the newly adapted models after model adaptation is performed in noisy speech recognition? In this case, we might have a clean speech model and estimated noise but no environment-matched noisy speech data because it is not always easy to collect enough noisy speech data to train the noisy model in advance. Then a new model can be adapted from clean one to match the testing environment by model adaptation techniques [4].

We found that the recognition performance of adapted model is still worse than the level obtained by the retrained model which is well trained by using noisy speech data [5]. It is interesting to investigate the possibility of improving the adapted performance up to the retrained one without noisy speech data. We attempt to provide an approximate error analysis based on an approximate measure on model difference using Kullback-Leibler (KL) relative divergence [6]. Then model separation can be performed by directly minimizing the KL relative divergence.

This paper is organized as follows. In section 2, the divergence-based model separation method is described. Experimental setup and results are provided in section 3. Final section gives the conclusions.

2. MODEL SEPARATION BASED ON KL DIVERGENCE

The proposed divergence-based model distance measure [6] is described as follows. The KL information (or KL
divergence) of a distribution \( q(x) \) with respect to another distribution \( p(x) \) has been defined as

\[
D_{KL}(p,q) = E_x \left[ \log \left( \frac{p(x)}{q(x)} \right) \right] = \sum_{x} p(x) \log \left( \frac{p(x)}{q(x)} \right)
\]  

(1)

where we define \( \log \frac{p}{q} = 0 \) and \( p \log \frac{p}{q} = \infty \). The KL information is greater than or equal to zero, and only equal to zero when the two distributions are identical. For a continuous density HMM (CDHMM), the mixture Gaussian density is widely used as state observation probability density function (pdf). We define an averaged KL information between the states as

\[
\bar{D}_{KL}(P,Q) = \frac{1}{M_p M_q} \sum_{i=0}^{M_p-1} \sum_{j=0}^{M_q-1} D_{KL}(p_{i,d},q_{j,d})
\]  

(2)

where \( M_p \) and \( M_q \) are the mixture numbers of the state \( P \) and \( Q \). The observation densities \( p_{i,d} \) and \( q_{j,d} \) are assumed to be mixture of Gaussians.

Next, we formulate a divergence-based misclassification function to measure the distance between two CDHMM

\[
R_{KL}(\Lambda_i,\Lambda_j) = \frac{1}{N} \sum_{m=1}^{N} \left[ D_{KL}(S_{m,i},S_{m,j}) + \left( \frac{1}{2} \sum_{d=1}^{D} D_{KL}(S_{m,i,d},S_{m,j,d}) \right) \right]
\]  

(3)

where \( N \) is the number of state of the models and \( S_{i,A} \) denotes the \( i^{th} \) state of the models \( \Lambda \). Note that the KL information is not symmetric. So, we define a symmetric relative divergence between two models as follows,

\[
\bar{R}_{KL}(\Lambda_1,\Lambda_2) = \frac{1}{2} R_{KL}(\Lambda_1,\Lambda_2) + R_{KL}(\Lambda_2,\Lambda_1)
\]  

(4)

The relative divergence, \( \bar{R}_{KL}(\Lambda_1,\Lambda_2) \), is to measure the difference between intra-state and inter-state distances for two models, \( \Lambda_1 \) and \( \Lambda_2 \). The smaller the relative divergence in Eq.(4) is, the more discriminative the two models are [6]. The discriminant functions, \( g(X,\Lambda_i) \), were estimated from the classification of input data \( X \) via HMM \( \Lambda_i \) in the priori techniques of discriminative training. We can directly obtained the discriminant function from the relative divergence, \( \bar{R}_{KL}(\Lambda_1,\Lambda_2) \), which is to approximate the recognition error rate. Thus, the training data is no more needed.

We expect the improvement in model discrimination based on relative divergence could provide some gain in accuracy since it is not theoretically justified. That is, the proposed method is to improve the discriminative powers for existing model which may be adapted (or transformed) from clean one. After model adaptation, the parameters of the model are supposed to match the testing environment. As mentioned earlier, the recognition performance of the newly adapted model is still worse than the level of retraining by using noisy speech data. Our approach is to improve the model discrimination in the model-space without any input speech since the existing model has already matched the testing environment.

The construction for adaptive training to minimize the relative divergence is as follows. The loss function can then be defined as

\[
\lambda(\Lambda_1,\Lambda_2) = \frac{1}{1+e^{-\gamma R_{KL}(\Lambda_1,\Lambda_2)+\theta}},
\]  

(5)

The parameters of speech models are adjusted in the model-space based on minimum relative divergence. The model parameters can be adaptively tuned by

\[
\Lambda_{1}^{(n+1)} = \Lambda_{1}^{(n)} - eU_n \nabla \lambda(\Lambda_{1}^{(n)},\Lambda_2),
\]

(6)

where \( n \) is the iteration and \( U_n \) is a positive definite matrix.

The discriminative adjustment of the mean vectors of HMMs follows

\[
\mu_{\omega_A,\Lambda_1}^{(n+1)} = \mu_{\omega_A,\Lambda_1}^{(n)} - e \frac{\partial \lambda(\Lambda_{1}^{(n)},\Lambda_2)}{\partial \mu_{\omega_A,\Lambda_1}},
\]

(7)

where

\[
\frac{\partial \lambda(\Lambda_{1}^{(n)},\Lambda_2)}{\partial \mu_{\omega_A,\Lambda_1}} = \frac{\partial \lambda(\Lambda_{1}^{(n)},\Lambda_2)}{\partial R_{KL}(\Lambda_{1}^{(n)},\Lambda_2)} \frac{\partial R_{KL}(\Lambda_{1}^{(n)},\Lambda_2)}{\partial \mu_{\omega_A,\Lambda_1}},
\]

(8)

\[
\frac{\partial \lambda(\Lambda_{1}^{(n)},\Lambda_2)}{\partial \mu_{\omega_A,\Lambda_1}} = eU_n \nabla \lambda(\Lambda_{1}^{(n)},\Lambda_2)(1 - \lambda(\Lambda_{1}^{(n)},\Lambda_2)),
\]

(9)

\[
\frac{\partial R_{KL}(\Lambda_{1}^{(n)},\Lambda_2)}{\partial \mu_{\omega_A,\Lambda_1}} = \frac{1}{2M_{\omega_A,\Lambda_1}} \sum_{n=1}^{M_{\omega_A,\Lambda_1}} \left[ \frac{\mu_{\omega_A,\Lambda_1} - \mu_{\omega_A,\Lambda_2}}{\sigma_{\omega_A,\Lambda_1}} \right] + \frac{1}{2M_{\omega_A,\Lambda_2}} \sum_{n=1}^{M_{\omega_A,\Lambda_2}} \left[ \frac{\mu_{\omega_A,\Lambda_2} - \mu_{\omega_A,\Lambda_1}}{\sigma_{\omega_A,\Lambda_2}} \right]
\]

(10)

and

\[
k_1 = \arg \min_{k \in \omega_A} D_{KL}(X_{\omega_A,k},S_{\omega_A,k}),
\]

(11)

\[
k_2 = \arg \min_{k \in \omega_A} D_{KL}(X_{\omega_A,k},S_{\omega_A,k}),
\]

(12)
and $M_{i,j}$ denotes the mixture number of $i^{th}$ state of model $\Lambda$, $\mu_{i,j}$ and $\sigma_{i,j}$ denote the standard deviation of $d^{th}$ order, $j^{th}$ mixture density, and $i^{th}$ state of model $\Lambda$, respectively.

The formula for adjustment of standard deviation follows

$$
\sigma^{(n+1)}_{i,j} = \sigma^{(n)}_{i,j} - e \frac{\partial \lambda(\Lambda^{(n)}_{i,j}, \Lambda)}{\partial \sigma_{i,j}},
$$

where

$$
\frac{\partial \lambda(\Lambda^{(n)}_{i,j}, \Lambda)}{\partial \sigma_{i,j}} = \frac{\partial \lambda(\Lambda^{(n)}_{i,j}, \Lambda)}{\partial \sigma_{i,j}} \frac{\partial \lambda(\Lambda^{(n)}_{i,j}, \Lambda)}{\partial \sigma_{i,j}} \frac{\partial \lambda(\Lambda^{(n)}_{i,j}, \Lambda)}{\partial \sigma_{i,j}} \frac{\partial \lambda(\Lambda^{(n)}_{i,j}, \Lambda)}{\partial \sigma_{i,j}}
$$

$$
+ \frac{\partial \lambda(\Lambda^{(n)}_{i,j}, \Lambda)}{\partial \sigma_{i,j}} \frac{\partial \lambda(\Lambda^{(n)}_{i,j}, \Lambda)}{\partial \sigma_{i,j}} \frac{\partial \lambda(\Lambda^{(n)}_{i,j}, \Lambda)}{\partial \sigma_{i,j}} \frac{\partial \lambda(\Lambda^{(n)}_{i,j}, \Lambda)}{\partial \sigma_{i,j}}
$$

and $M = 1, 2$ mixture density, and $\sigma^2_s$, $\sigma^2_a$ denote the variance of testing speech utterance and variance of added noise. Traditional parallel model combination (PMC) [10] was used to adapt clean model to match the testing environment, where only the parameters of static mean vectors and covariance matrices were adapted. The model separation experiments were performed in state and sub-syllable levels, it implies that only some parameters will be tuned according to a pre-defined condition, where the condition is defined by some divergence thresholds or by the topology of speech models.

In the application of improving model discrimination, we set $\Lambda = \Lambda = \Lambda$ in above equations. The model separation training can be performed in state or subword level. It implies that only some parameters will be tuned according to a pre-defined condition, where the condition is defined by some divergence thresholds or by the topology of speech models.

### 3. EXPERIMENTAL RESULTS

There are at least two approached to model the sub-syllable units of Mandarin. Traditional one is so-called initial/final structure [7]; there are 22 initials and 38 finals. We use a preme/core-final structure to model the half-syllable units of Mandarin [8]; there are 61 premes and 25 core finals. The main difference between these two approaches is the modeling of the “glide”, the final containing a medial $i$, $u$ or $y$. In preme/core-final structure, the medial is split from the final. Then, the set of finals is reduced to 25 core-finals while the set of premes is extended by legally combining an initial with a medial.

The evaluation experiments were performed on the telephone corpus MAT (Mandarin Across Taiwan) [9]. The corpus MAT-560 including MAT-160 and MAT-400 was used for training, and the testing set includes 500 utterances which were used for self-evaluation in Taiwan Mandarin benchmark. The sub-syllable units in the preme/core-final structure are modeled by continuous density HMM with diagonal covariance matrices. Front-end feature consists of 12-order Mel frequency cepstral coefficients (MFCC) and its first-order time derivatives. Sentence based cepstrum mean normalization (CMN) was applied. Free syllable decoding without any treatment on tone was performed for performance evaluation.

White Gaussian noise was added to the testing utterances at different SNR levels. The sentence-based SNR is defined as follows,

$$
SNR = 10 \log_{10} \left( \frac{\sigma^2_s}{\sigma^2_a} \right) (dB)
$$

where $\sigma^2_s$ is the variance of testing speech utterance and $\sigma^2_a$ is the variance of added noise. Traditional parallel model combination (PMC) [10] was used to adapt clean model to match the testing environment, where only the parameters of static mean vectors and covariance matrices were adapted. The model separation experiments were performed in state and sub-syllable levels, it implies that only some parameters will be tuned according to a pre-defined condition, where the condition is defined by some divergence thresholds or by the topology of speech models.

The experimental results are summarized in Table 1. The experiments were performed at different SNR levels of 30, 20, 10, and 0 dB. The experimental results in Table 1 are shown in syllable accuracy obtained from free syllable decoding.

<table>
<thead>
<tr>
<th>Method</th>
<th>Clean 30dB</th>
<th>20dB</th>
<th>10dB</th>
<th>0 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>50.7</td>
<td>47.9</td>
<td>37.2</td>
<td>18.4</td>
</tr>
<tr>
<td>Adapted (PMC)</td>
<td>50.1</td>
<td>43.4</td>
<td>29.1</td>
<td>10.6</td>
</tr>
<tr>
<td>State MS</td>
<td>50.5</td>
<td>50.2</td>
<td>44.0</td>
<td>29.6</td>
</tr>
<tr>
<td>Sub-syllable MS</td>
<td>50.6</td>
<td>50.4</td>
<td>44.3</td>
<td>30.0</td>
</tr>
</tbody>
</table>

Table 1. Syllable accuracy (%) of different methods, where MS means model separation.

It shows that there is no gain in accuracy for clean condition by using state-based or sub-syllable-based model separation. The reason might be that the clean model has been well trained. We also found that the syllable-based model separation performed slight better than state-based one for all different SNR levels. It did improve the accuracy for the adapted models although it has not been proven it from a theoretical point of view. Note that the experimental results of model separation are only obtained from tuning the mean vectors of HMMs. In our experiments, to adjust the variance vectors in the
proposed algorithm didn’t achieve any improvement in recognition accuracy.

4. CONCLUSIONS

In this paper, we have reported the primary results of divergence-based model separation. The application could be to enhance the model discrimination after some existing algorithms transform speech model to match the testing environment. Unlike traditional training schemes, the training data might be not available. Although the proposed approach has not been theoretically proven, experimental results show that some improvements in accuracy could be obtained by the proposed algorithm. It’s still a very interesting issue to enhance the discriminative power of adapted models while the environment-matched data is not available in robust speech recognition.

5. ACKNOWLEDGEMENT

The first author would like to express his sincere gratitude to Dr. Chin-Hui LEE for his valuable comments and discussions on this framework.

6. REFERENCES