Minimum Phone Error Discriminative Training For Mandarin Chinese Speaker Adaptation

Liang-Yu Chen [1], *Chun-Jen Lee* [2], *Jyh-Shing Roger Jang* [1]

[1] Department of Computer Science, National Tsing Hua University, Taiwan
[2] Telecommunication Labs., Chunghwa Telecom Co., Ltd., Taiwan
davidson833@mirlab.org, cjlee@cht.com.tw, jang@cs.nthu.edu.tw

Abstract

Speaker adaptation is an efficient way to model a new speaker from an existing speaker-independent model with limited speaker-dependent data. In this paper, we investigate the use of discriminative training schemes based on the minimum phone error (MPE) criterion to improve a well-known speaker adaptation technique, a combination of transform-based adaptation and Bayesian adaptation. Furthermore, a new approach utilizing the statistics of the model-based regression tree for controlling the interpolation between maximum likelihood (ML) and MPE objective functions is also presented. Several comparative experiments were conducted on a continuous speech recognition task for Mandarin Chinese. Experimental results show that the proposed approach can further improve the performance of the original hybrid adaptation.

Index Terms: speaker adaptation, discriminative training, minimum phone error

1. Introduction

The performance of speech recognition degrades rapidly when there is a mismatch between the test and the training conditions, such as a mismatch between speakers. One practical approach to solve this problem is to adapt an existing speaker-independent (SI) model to a speaker-dependent (SD) model with some speaker-specific data. In general, there are two typical approaches for speaker adaptation: one is Bayesian adaptation, where the acoustic model parameters are directly adjusted based on a Bayesian framework, such as maximum a posteriori (MAP) adaptation, whereas the other is transform-based adaptation, where clusters of model parameters are transformed individually based on cluster-specific transform functions, such as maximum likelihood linear regression (MLLR). Relevant studies revealed that a hybrid approach, combining MAP adaptation and transform-based adaptation, have been convinced to be better than MAP or transform-based adaptation alone [1].

Recently discriminative training criteria have been widely employed to estimate more accurate HMM models for speech recognition, such as maximum mutual information (MMI) [2], minimum classification error (MCE) [3], and minimum phone error (MPE) training [4]. Therefore, several studies focusing on discriminative speaker adaptation have been reported. Uebel and Woodland [5] applied an interpolation of ML and MMI training criteria to estimate discriminative linear transform. Wang and Woodland [6] adopted the MPE criterion to estimate discriminative linear transform. In addition, Povey et al. [7] investigated the integration the MAP scheme into MMI and MPE for task and gender adaptation, respectively.

Based on the observations above, we are motivated to present a framework which combines discriminative training and speaker adaptation for robust speech recognition. Furthermore, an alternative approach for setting the interpolation values between ML and MPE objective functions based on the statistics of the regression tree is also proposed. Experiments on Mandarin Chinese speaker adaptation are conducted to illustrate the improvement of the proposed approach.

The rest of the paper is organized as follows. Section 2 introduces our MPE based discriminative training for speaker adaptation; Section 3 proposes a regression tree based criterion for setting the interpolation factor. The experimental setup and a quantitative assessment of achieved performance are presented in Section 4, and Section 5 concludes our findings.

2. Discriminative training for speaker adaptation

In this paper, we focus on discriminative training based on the MPE criterion for speaker adaptation. A two-stage approach is investigated to achieve the goal of discriminative speaker adaptation.

2.1. A hybrid adaptation (MLLR+MAP)

In the first stage, a hybrid adaptation is adopted which employs MLLR followed by MAP adaptation, denoted by "MLLR+MAP". Based on this approach, the estimated model mean for the mixture component *m* of state *j* is expressed as follows:

$$\hat{\mu}_{jm} = \frac{N_{jm}}{N_{jm}+\tau}\tilde{\mu}_{jm} + \frac{\tau}{N_{jm}+\tau}\mu_{jm} \qquad (1)$$

where μ_{jm} is the speaker-dependent mean transformed by MLLR, $\tilde{\mu}_{jm}$ is the mean of the observed adaptation data, τ is a weighting factor of the *a priori* knowledge to the adaptation data, and N_{jm} is the occupation probability of the adaptation data.

Applying speaker adaptation to acoustic models generally improves the total recognition rate. However, fine adjustment for each acoustic HMM model is not guaranteed by using such approach alone. To make the matter worse, the adjustment of HMM parameters by adaptation may trigger some acoustic models to blend with their surroundings more closely. This is especially the case for those HMM models that are confused in the training data. Figure 1 depicts some statistics of phone recognition errors (substitution errors) after applying MLLR+MAP. Each line in Fig. 1 represents a phone model being adapted with different amount of adaptation data. As we can see, the performance of some phone models, with high substitution errors, deteriorates as the amount of adaptation sentences increases. Figure 1 also shows the need to alleviate the confusion between competitive phone models via a discriminative processing.

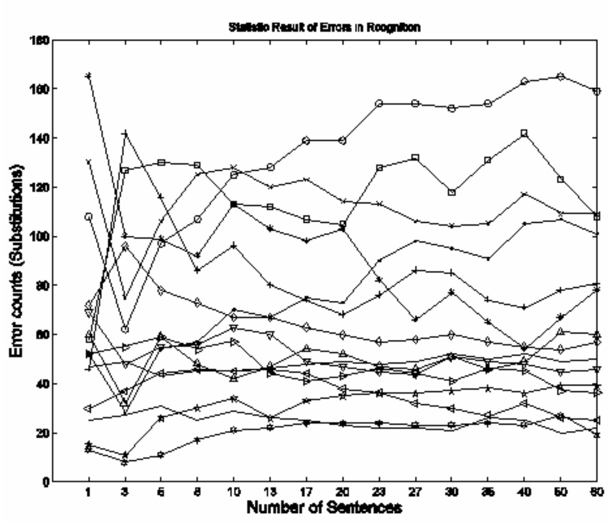

Figure 1: *Statistics of phone recognition errors (substitution) after applying MLLR+MAP.*

2.2. Integrated approach (MLLR+MAP+MPE)

In the second stage, a MPE discriminative training is applied based on speaker-specific data.

By using a weak-sense auxiliary function in HMM estimation, the mean $\tilde{\mu}_{jm}$ and the variance $\tilde{\sigma}^2_{jm}$ of mixture component *m* of state *j* of a new HMM parameter set $\tilde{\lambda}$ can be re-estimated as follows:

$$\tilde{\mu}_{jm} = \frac{\{\theta^{num}_{jm}(O) - \theta^{den}_{jm}(O)\} + D_{jm}\mu_{jm}}{\{\gamma^{num}_{jm} - \gamma^{den}_{jm}\} + D_{jm}} \tag{2}$$

$$\tilde{\sigma}^2_{jm} = \frac{\{\theta^{num}_{jm}(O^2) - \theta^{den}_{jm}(O^2)\} + D_{jm}(\sigma^2_{jm} + \mu^2_{jm})}{\{\gamma^{num}_{jm} - \gamma^{den}_{jm}\} + D_{jm}} - \tilde{\mu}^2_{jm} \tag{3}$$

where D_{jm} is the Gaussian-specific smoothing constant, and $\theta_{jm}(O)$ is the sum of observation data weighted by the occupation probability for mixture component *m* of state *j*, $\theta_{jm}(O^2)$ is the sum of squared observation data weighted by the occupation probability for mixture component *m* of state *j*; γ^{num}_{jm} and γ^{den}_{jm} are the numerator occupation probabilities and the denominator occupation probabilities summed over time respectively.

Besides, as mentioned in [4], I-smoothing can be interpreted as an interpolation between ML and MPE objective functions, formulated as:

$$\gamma^{num}_{jm}\,' = \gamma^{num}_{jm} + \tau_{jm} \tag{4}$$

$$\theta^{num}_{jm}(O)' = \theta^{num}_{jm}(O) + \frac{\tau_{jm}}{\gamma^{mle}_{jm}}\theta^{mle}_{jm}(O) \tag{5}$$

$$\theta^{num}_{jm}(O^2)' = \theta^{num}_{jm}(O^2) + \frac{\tau_{jm}}{\gamma^{mle}_{jm}}\theta^{mle}_{jm}(O^2) \tag{6}$$

where the superscript *mle* indicates occupation probabilities obtained from the alignment of the transcriptions by ML, and τ_{jm} represents a weighting factor for the contribution of ML. The estimation of the parameters in (2) and (3) is then obtained by using (4)–(6).

According to section 2.1, the hybrid approach to speaker adaptation is based on MLLR transformation followed by MAP adaptation. Therefore, occupancies obtained with ML prior, in (5) and (6), would be substituted by the speaker-adapted models, i.e. the MLLR+MAP prior.

Our approach is motivated by considering the use of MPE training to fine-tune the SD acoustic models. The proposed approach of discriminative training for speaker adaptation is illustrated with the flow chart shown in Figure 2. Briefly, the SI model is first adjusted by MLLR followed by MAP adaptation with limited speaker-specific data. Then, the adapted SD model is further updated by a lattice-based MPE training. The numerator lattice is obtained by the forced alignment process on the correct transcriptions of the adaptation data, while the denominator lattice is approximated with n-best phone string hypotheses via the recognition process on the adaptation data [4], [8].

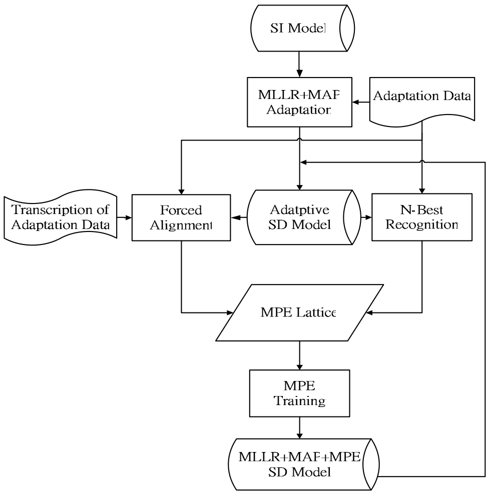

Figure 2: *Schematic diagram of discriminative training for speaker adaptation*

3. Regression tree based MPE (RTMPE)

3.1. Regression tree construction

To discriminate possible confusion phones, we propose a new approach based on the statistics of the regression tree for controlling the interpolation between ML and MPE objective functions to discriminate possible confusion phones mentioned in Section 2.1.

For a given set X of mixture components, the Bayesian information criterion (BIC) [9] is applied to split HMM model parameters into clusters, formulated as:

$$BIC(M_i, X) = \log p(X \mid M_i) - \frac{1}{2}\#(M_i)\log n \quad (7)$$

where $p(X \mid M_i)$ is the likelihood of the set X for the HMM model M_i, $\#(M_i)$ is the number of parameters of M_i, and n is the number of components of X. This $BIC(M_i, \mathrm{X})$ value represents the likelihood of modeling data X with model M_i, while taking the consideration of the number of parameters used in the model as a penalty.

Regression tree construction is performed by the top-down approach: parent clusters are split iteratively into finer children clusters until the stopping criterion is reached. Herein, a BIC-based stopping criterion for a given cluster C_i is applied and formulated as:

$$\Delta BIC_{21}(C_i) = BIC(GMM_2, X) - BIC(GMM_1, X) \quad (8)$$

where GMM_j is a Gaussian mixture model with j mixture components. When the $\Delta BIC_{21}(C_i)$ value is positive, i.e. modeling C_i with two clusters (two Gaussian mixture models) has higher probability than modeling it with only one cluster, the node C_i will be split. When the $\Delta BIC_{21}(C_i)$ value is negative, the node C_i will not be split.

The algorithm for constructing the regression tree is described as follows:

1. Initially group all HMM mixture components into one cluster, i.e. the root node of the regression tree (RT).
2. For a given cluster C_i (leaf node) of RT, compute $\Delta BIC_{21}(C_i)$.
3. Split C_i into two new clusters (children leaf nodes) if C_i has sufficient amount of component ($n > \delta_0$) and $\Delta BIC_{21}(C_i) > 0$.
4. Repeat steps 2 and 3 until no leaf node in RT can be split.

In the step 3 of the above algorithm, n is the number of components in C_j and δ_0 is a threshold for controlling the minimum amount of data in a node. This makes sure that a leaf node with insufficient amount of data will not be split and the time to construct the regression tree can be reduced.

3.2. Selection of Gaussian mixture component

The tree construction process reflects that a node in the lower level of the regression tree consists of closer mixture components. To control the setting of I-smoothing for different Gaussian mixture components based on the regression tree RT constructed above, the representative mixture components scattered in the leaf nodes of RT are further explored. The algorithm for selecting representative mixture components is described as follows:

1. For a given cluster C_i (leaf node) containing different phone models $\{p_k\}$, the following conditions are verified:
   (1) Check if the number of mixture components in p_k in C_i divided by the total number of mixture components in C_i is above the threshold δ_1.
   (2) Check if the number of mixture components in p_k in C_i divided by the total number of mixture components in p_k in RT is above the threshold δ_2.
   (3) Check if there are at least two or more different phone models $\{p_k\}$ in C_i satisfying both condition (1) and condition (2).
2. Add those $\{p_k\}$ which satisfy the above conditions into a set called the Dominant Set (DS).
3. Repeat steps 1 and 2 until all leaf nodes in RT have been checked.

The mixture components in DS represent the dominant components in each cluster. It implies that less weighting should be contributed by ML training and more weighting should be contributed by MPE training for estimating HMM model parameters for the mixture components in DS. The approach of I-smoothing setting for MPE training is denoted by "Regression Tree based MPE" (RTMPE). The effect of the proposed discriminative training for speaker adaptation with RTMPE will be presented in the following section.

4. Experiments

4.1. Experimental setup

Several experiments of adaptation for Mandarin Chinese speakers were conducted for comparison. Two microphone databases were collected and down-sampled to 16 kHz for training SI and SD HMM models, respectively. The SI database, consisting of about 9500 short sentences, was recorded by 100 male and 100 female speakers. The SD database was recorded by 11 male and 10 female speakers to evaluate our adaptation methods. In the SD database, 160 short sentences were recorded by each speaker, with 10 to 15 Chinese characters per sentence, where 60 sentences are used for adaptation and 100 sentences for testing. All experiments were carried out in supervised mode.

Mandarin Chinese is a syllable-based language. A syllable is composed of a syllable initial followed by a syllable final. Since the coarticulation effects within a syllable are more significant than those across syllables, only intra-syllable context modeling is used in this study. Both initial and final units were modeled by 3-state left-to-right HMMs with no skipped states, and the silence unit was modeled by a 1-state HMM. The HMM models consist of about 2200 Gaussian mixtures in total. The acoustic analysis is performed at 10 ms frame rate using a 20 ms hamming window. Each frame contains 24 spectral feature coefficients, including 12 MFCC and their delta values.

4.2. Experimental results

Figure 3 and Figure 4 illustrate the sensitivity of phone error rates of SI, MLLR+MAP, and MLLR+MAP+MPE to the number of the adaptation sentences in the cases of male and female speakers respectively. In comparison with SI baselines, applying MLLR+MAP adaptation could reduce the phone error rates from 31.77% to 23.76% for male speakers and from 26.50% to 16.49% for female speakers, for the case of using 60 adaptation sentences. It is also observed that adaptation with MPE training, MLLR+MAP+MPE, can generally improve the adaptation performance for both male and female speakers.

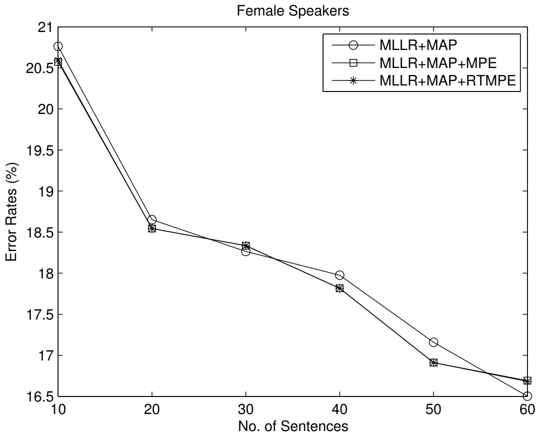

Figure 3: Comparison of MLLR+MAP and MLLR+MAP+ MPE for female speakers; the SI baseline is 26.50%.

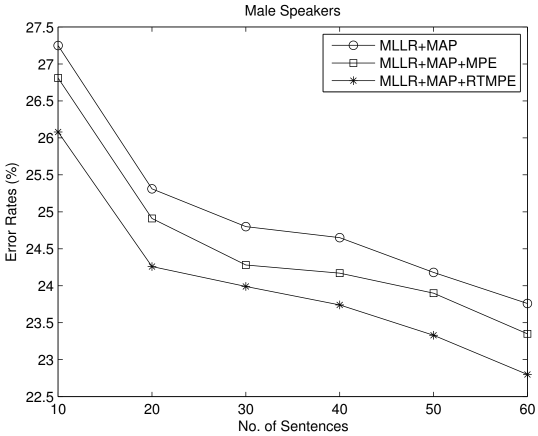

Figure 4: *Comparison of MLLR+MAP, MLLR+MAP+MPE, and MLLR+MAP+RTMPE for male speakers; the SI baseline is 31.77%.*

The next experiment is to use the RTMPE approach after MLLR+MAP. Currently, we simply set δ_0, δ_1, and δ_2 to the constant values 80, 1/3, and 1/3, respectively. The performance of MLLR+MAP+RTMPE in terms of phone error rates for both female and male speakers is also shown in Figure 3 and Figure 4. From Figure 4, the discriminative training approaches, MLLR+MAP+MPE and MLLR+MAP+RTMPE, achieve better performance than the approach with the hybrid adaptation alone, namely MLLR+MAP. Furthermore, the results illustrate that the proposed two-stage approach with MLLR+MAP+RTMPE achieves the best performance than the others, giving a total of 1% improvement to the original MLLR+MAP approach. It also reveals that special concern on I-smoothing for distinctive mixture components based on the statistics of the regression tree is pertinent to MPE training, although the improvement is slight. The female results show no improvement from MLLR+MAP+RTMPE, giving almost the same performance as MLLR+MAP+MPE. This might be the result of lower error rates for the female speakers so that the discriminative training does not render much effect as for the male speakers.

5. Conclusions

In this study we have presented an approach based on the MPE criterion to Mandarin Chinese speaker adaptation. In addition, a method for setting the values of interpolation factors to different mixture components according to the statistics of the regression tree has also been investigated. Experimental results have shown that the performance of the proposed discriminative schemes can be improved continuously and consistently as the number of adaptation data increases. As compared to the well-known hybrid approach which combines MLLR and MAP adaptation, the proposed approach with discriminative training gives better performance.

6. Acknowledgement

We would like to thank the other members of the MIR Laboratory, National Tsing Hua University, for their help with setting up part of the experiments, in particular Jiang-Chun Chen and Hsueh-Wen Niu.

7. References

[1] Digalakis, V. V. and Neumeyer, L. G., “Speaker adaptation using combined transformation and Bayesian methods,” *IEEE Trans. Speech Audio Processing*, vol.4, pp. 294-300, July 1996.

[2] Woodland, P. C. and Povey, D., “Large scale discriminative training of hidden Markov models for speech recognition,” *Computer Speech and Language*, vol. 6, pp. 25-47, 2002.

[3] Juang, B.-H., Chou, W. and Lee, C.-H., “Minimum classification error rate methods for speech recognition,” *IEEE Trans. Speech Audio Processing*, vol. 5, no. 3, 1997, pp. 257-265.

[4] Povey, D., “Discriminative training for large vocabulary speech recognition,” *Ph.D Dissertation*, *Department of Engineering*, *University of Cambridge*, July 2004.

[5] Uebel, L. F. and Woodland, P. C., “Discriminative linear transforms for speaker adaptation,” *Proc. ISCA ITRW on Adaptation Methods for Automatic Speech Recognition*, Sophia-Antipolis, 2001.

[6] Wang, L. and Woodland, P. C., “MPE-based discriminative linear transform for speaker adaptation,” *Proc. ICASSP*, pp. 321-324, 2004.

[7] Povey, D., Gales, M.J.F., Kim, D.Y. and Woodland, P. C., “MMI-MAP and MPE-MAP for acoustic model adaptation,” *Proc. Eurospeech*, pp. 1891-1894, 2003.

[8] Chen, J.-C., Lee, C.-J., Hsu, S.-P. and Jang, J.-S. R., “Sausage-net-based minimum phone error training for continuous phone recognition,” *Proc. ISCSLP*, vol. 2, pp. 280-288, 2006.

[9] Fraley, C. and Raftery, A. E., “How many clusters? Which clustering method? Answers via model-based cluster analysis,” *Computer Journal*, vol. 41, pp. 578-588, 1998.