MINIMUM PHONE ERROR MODEL TRAINING ON MERGED ACOUSTIC UNITS FOR TRANSCRIBING BILINGUAL CODE-SWITCHED SPEECH

Ching-Feng Yeh[1],Yiu-Chang Lin[2], Lin-Shan Lee[1,2]

Graduate Institute of Communication Engineering, National Taiwan University, Taiwan[1]
Department of Electrical Engineering, National Taiwan University, Taiwan[2]
andrew.yeh.1987@gmail.com

ABSTRACT

This paper proposes to perform Minimum Phone Error (MPE) model training on merged acoustic units for transcribing Mandarin-English code-switched lectures with highly imbalanced language distribution. Some of the acoustic events in Mandarin and English may have very similar characteristics, so the states or Gaussian mixtures representing them can be merged with identical shared parameters. When MPE is performed afterwards, these merged identical states or Gaussian mixtures can form a compact acoustic unit set. In this way MPE can better discriminate the acoustic units of both languages, because similar units are merged while distinct units are differentiated. Significant improvements in recognition accuracy were observed in the preliminary experiments on real-world bilingual code-switched lecture corpus recorded at National Taiwan University.

Index Terms— MPE, bilingual, merging, discriminative, code-switching

1. INTRODUCTION

In the globalized world today, many people are using more than a single language in their daily lives. As a result, very often the speech signals include more than one language. This is why great effort has been made to try to extend existing speech recognition technologies primarily developed for monolingual tasks to consider multilingual environments [1][2]. A major concern here is the phoneme sets of different languages. Very often some phonemes are shared by different languages; some phonemes in different languages sound similar, but slightly different; and some other phonemes are unique for specific languages. This makes acoustic modeling and lexicon construction difficult. Usually the similarity and differences between the phonemes and the unique characteristics for some phonemes are difficult to measure quantitatively. Many approaches have been proposed to merge acoustic units on different levels to handle these problems [3][4][5][6][7].

In general, bilingual speech can be classified into two categories. The first one is inter-sentential switching, in which the speaker switches languages from sentences to sentences. For example, the sentence, "It's fine. 謝謝你(Thank you).", where the first sentence is in English, while the second in Chinese. The other is intra-sentential switching, also referred to as code-switching in this paper, in which the languages are switched from words to words. For example, the sentence "這個 equation 很複雜. (This equation is very complicated.)", in which the word "equation" in the guest language of English is embedded in a sentence in the host language of Chinese. The latter case of code-switching is very common for speakers with non-English native language (with English as the guest language and the non-English native language as the host), especially when they speak very good English and many English words are not yet properly translated into their native languages.

An extra difficulty for the above second category of code-switched bilingual speech is the highly imbalanced language distribution, i.e., in such cases there are much more host language data but very limited guest language data. This makes not only acoustic modeling for the guest language difficult, but the recognizer usually tends to take most speech as in the host language. This second category of code-switched bilingual speech is the target of this paper. Another distinguishing feature for the above second category of code-switching environment is the issue of language identification [8][9]. Since the language may be switched back and forth from word to word within an utterance and this makes the language identification harder [9]. In addition, in many cases of such code-switched bilingual speech, it may be more important to correctly transcribe the words in the guest language than those in host language, very often those words in the guest language are the key terms for speech understanding.

In this paper, we applied minimum phone error (MPE) [10] training technique to code-switched speech mentioned above. Similar results were reported earlier [11], in which it was found that MPE brought relatively limited improvement to code-switched speech. In this paper, we present possible reasons for the limited improvement and propose to apply MPE on acoustic units merged on the state and Gaussian levels

2. CODE-SWITCHED TESTING ENVIRONMENT

The corpus used for the experiments reported here was the recorded lecture of courses offered in National Taiwan University. The speech is very spontaneous, belonging to the

above second category of code-switching, i.e., in host language of Mandarin and guest language of English, with highly imbalanced Mandarin / English percentage ratio. The detailed description of experimental data is listed in Table 1. We see in the last row of this table the percentage of English (guest language) for this bilingual corpus is only 15.2%, or roughly 1.5 hours in the training set. Such a highly imbalanced data distribution makes the recognition task very difficult, especially the accuracy for the guest language words turn out to be very low, as will be shown below.

Content	**Course Lectures**
Speaking Style	*Spontaneous Monologue*
Training Set (hr)	9.10
Development Set (min)	126.81
Testing Set (min)	133.77
Mandarin / English (%)	84.8 / 15.2

Table 1. *Details for the Target Corpora*.

3. BASELINE EXPERIMENT ON MINIMUM PHONE ERROR (MPE) TRAINING ON THE TARGET CORPUS

Maximum likelihood (ML) criterion was conventionally adopted as a common approach for estimating acoustic model parameters. Discriminative training approaches such as Minimum Error (MPE) [10] training has been popularly used in recent years to discriminate the signal feature distributions for similar acoustic units. MPE training takes the competing candidates in the decoded lattices as additional information into consideration and tries to distinguish these competing candidates by adjusting the corresponding model parameters. This is different from the ML criterion, which only tries to maximize the likelihood for each respective candidate without considering the competing models.

In general, MPE outperformed ML in most speech recognition systems in the most cases. However, this is not necessarily true for the code-switched bilingual speech task considered here. Because of the high degree of ambiguity between similar acoustic units for the two different languages, MPE may actually over-discriminate some very similar acoustic units because these units are labeled as different in different languages. For example, the plosive /b/ in Mandarin (/CH_b/) and in English (/EN_B/) are very similar, especially in the code-switched corpus considered they were produced by the same speaker. However, the standard MPE training procedure tries to differentiate the models for these two phonemes by increasing the distance between them since they are differently labeled in different languages. This may be the reason why MPE did not bring expected recognition performance improvement for multilingual tasks [11] as usually obtained in monolingual task. This is also consistent with the observation in the testing environment as described in Table 1. The results for the baseline experiment are shown in Fig. 1.

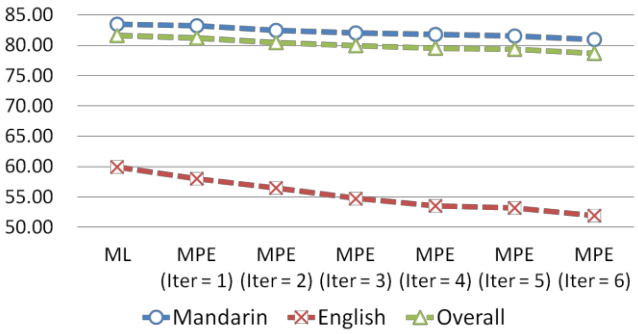

Figure 1: *ML / MPE Results for the Target Corpora (Acc.)*

Word accuracy for English, character accuracy for Mandarin and summation for overall were shown in Fig. 1. In Fig. 1, the recognition accuracy for English part is much lower in both ML and MPE case, obviously due to the highly imbalanced nature of code-switched bilingual corpus. More importantly, applying MPE training did not bring any improvement but instead the performance degraded monotonically as more iteration was performed. The degradation for English part is much faster than Mandarin, probably also due to the data-imbalanced nature. As mentioned above, English words here are usually key words, therefore it is definitely an important issue here.

4. PROPOSED APPRAOCHES

To handle the problem mentioned above, we proposed to merge similar acoustic units on state and Gaussian levels so they share the same parameters and become identical. MPE is then applied to differentiate the distinct acoustic units.

4.1. Complete System Architecture

The complete system flow chart for the proposed approach of MPE training on merged acoustic units is in Fig. 2.

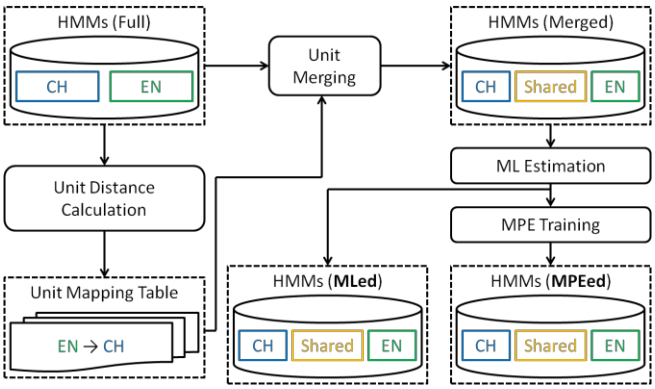

Figure 2: *Proposed System Architecture*

In Fig. 2, the baseline acoustic model set trained with bilingual corpus with the original bilingual phoneme set, used as the initial model, is at the upper left corner labeled as "HMMs (Full)". Distances between model units on two different levels, state or Gaussian, are then calculated for model units in "HMMs (Full)" and a mapping table is obtained,

as in the lower left corner of Fig. 2. In this work, since English is the guest language, for every English unit, the best mapped Mandarin unit with minimum distance is obtained. It is possible that many English model units are mapped to the same Mandarin model unit.

It is certainly reasonable to merge directly the acoustic models for English and Mandarin on the phoneme level. However, there are some constraints that limit the performance. Considering the linguistic nature of the two languages considered here, Mandarin is a tonal language while English is a stress-timed language. The different linguistic natures make it much more difficult to merge English and Mandarin acoustic units on the phoneme level (triphones) than to merge them on the lower levels of states or Gaussians. In fact, it has been found previously that merging acoustic units on the state and Gaussian levels offered much better improvements than merging on the phoneme level for the Mandarin / English task [3][4][6].

The next step is to perform acoustic unit merging on the set "HMM (Full)" based on the acquired mapping table. This merging process is started from unit pairs with minimum distance and repeated until desired number of merged English units is reached. The model acquired is at the upper right corner of Fig. 2, labeled as "HMMs (Merged)". To reach a better likelihood for the new model configuration for the training data, another ML estimation process for the merged model is performed, as shown below "HMMs (Merged)" in Fig. 2. This ML process produces a set of acoustic models labeled as "HMMs (MLed)" shown at the lower middle of Fig. 2. Although MPE was not performed yet, sharing of training data for the acoustic units is helpful for better modeling of the English model units. Therefore, better recognition accuracy is expected with this set of models as well [3]. Now because the similar model units (Gaussians or states) have been merged and made identical, the ambiguity between distinct acoustic units is largely reduced. So we applied MPE training to this model set, and the model set obtained after MPE training, labeled as "HMMs (MPEed)" at the lower right corner of Fig. 2, is to be used for transcribing the bilingual speech.

4.2. Kullback–Leibler Divergence

KL Divergence is popularly used as estimation between distances between two stochastic distributions. The KL Divergence of two Gaussian distributions is,

$$KLDiv(P_i, P_j) = \frac{1}{2}[\, ln\left(\frac{det\ (\Lambda_j)}{det\ (\Lambda_i)}\right) + trace(\Lambda_j^{-1}\Lambda_i) + (\mu_j - \mu_i)^T \Lambda_j^{-1}(\mu_j - \mu_i) - N\,], \quad (1)$$

where $\mu_{i,j}$ and $\Lambda_{i,j}$ are mean vector and covariance matrix of the Gaussian distributions $P_{i,j}$ and N is the dimensionality of x. Since KL Divergence is asymmetric, an alternative symmetric distance is defined and used in this work,

$$D_{KL}(P_i, P_j) = KLDiv(P_i, P_j) + KLDiv(P_j, P_i)\,. \quad (2)$$

4.3. State Level Distance

The state of HMMs can be considered as sequential components of the phonemes, each with a relatively steady feature distribution, which can be physically interpreted as the feature distribution produced by a certain acoustic event due to a distinct vocal tract shape. Since speech production by physical structure of vocal tract is universal across different languages, state may be a good unit to analyze cross-lingual similarity for acoustic modeling. Because the state duration is usually very short and not identifiable by human perception, in this work, we first model each state by a single Gaussian distribution, and then calculate the distance between states using the symmetric KL Divergence based on these single Gaussians as in (2),

$$D_S(S_i, S_j) = D_{KL}(G_i, G_j)\,, \quad (3)$$

where G_i is the single Gaussian representing state S_i. This approach has been shown to be useful in previous work [3][4].

3.3. Gaussian Level Distance

Since there can be many Gaussians in a state, Gaussian units describe the finer structure of the state units. Previous work has shown that Gaussian level mapping is very useful for identifying similar model units [3][7]. The calculation of distance between two Gaussians is simply based on the symmetric KL Divergence formula in (2),

$$D_G(G_i, G_j) = D_{KL}(G_i, G_j)\,. \quad (4)$$

However, the physical interpretation of Gaussian units is much weaker. The similarity between two Gaussians of two states does not imply the similarity between two states, since a Gaussian is only a fine structure component of the feature distribution of a state. However, merging similar Gaussians makes sense because the Gaussians can be enhanced by sharing the data, and similar Gaussians can be made identical and not differentiated by MPE.

5. EXPERIMENT

5.1. Experiment Setup

The corpus used for experiments here are reported in section 2.1 and listed in Table 1. The acoustic models used are in the format of state-tied triphone models. The bilingual lexicon used here included English words, Chinese words and all commonly used Chinese characters. Target-domain related corpora including frequency counts were used for both English and Chinese word selection for the lexicon. Chinese words were also generated by segmenting a large corpus using PAT-Tree base approaches [3]. We used the Kneser-Ney tri-gram model, started with a background model and then adapted with the transcription of the training set for the target lecture here.

The way the recognition accuracy was evaluated followed the earlier works [3], [12]. That is, when aligning recogni-

tion results with the reference transcriptions, insertions, deletions, substitutions were evaluated respectively for each language and summed up for overall evaluation. The basic unit for alignment is character for Mandarin and word for English [3][12], so the accuracies reported here are with respect to characters for Mandarin and to words for English.

5.2. Experimental Results

We conducted experiments for both ML estimation and MPE training by merging acoustic units on states and Gaussians for different percentages of English units being merged with Mandarin units. The experimental results for state-level merging are shown in Fig. 4. The vertical axis is the recognition accuracy while the horizontal axis represents the percentage of English state being merged. The red curve is for English words using ML estimation only. We can see that with more English states being merged, the accuracy for English words with ML estimation is improved continuously in general. Clearly this improvement was brought by data-sharing, which makes it possible to use Mandarin data and English data jointly to estimate the state parameters. The orange curve is then the result for MPE training with state merging. It is clear that the accuracy also increase when more English units were merged. More importantly, now the MPE results obviously outperformed ML results with state merging just as the monolingual case in general. This is quite different from those seen in Fig. 1, where MPE training in fact degraded the accuracy. This is achieved by unit merging. For Mandarin part, on the other hand, MPE offered only very limited improvement over ML, and both ML and MPE cases are only slightly influenced by state merging. It seems that MPE primarily offered improvements for the weak or confusing models, while in the case here most weak or confusing models are of English. This may be the reason why English was improved but Mandarin remained almost unchanged.

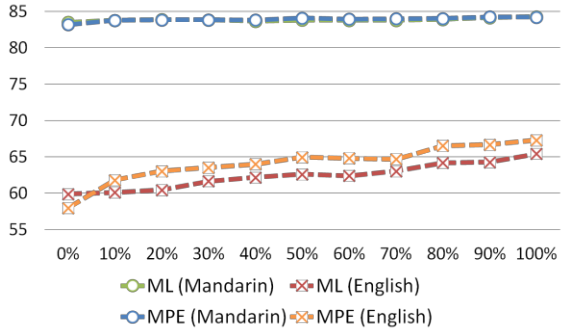

Figure 3: *ML and MPE Results by State Merging (Acc.)*

For Gaussian-level merging, the corresponding results are shown in Fig. 4. Very similar situation as in Fig. 3 can be observed. Therefore, the ambiguity in bilingual acoustic units is reduced by unit merging, on either state or Gaussian level, which makes it possible for MPE to offer improvements in accuracy. In the results shown in Fig. 3 and 4, only one iteration was performed for MPE. Better improvements may be possible with more iteration.

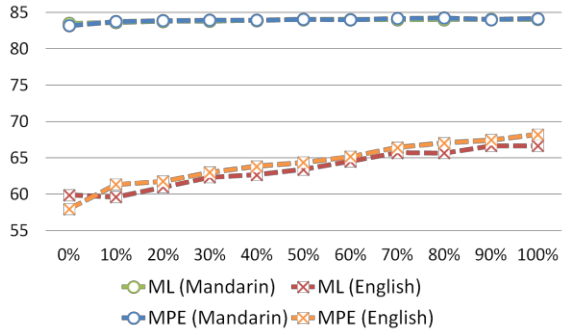

Figure 4: *ML and MPE Results by Gaussian Merging (Acc.)*

Table 2 gives the detailed accuracies. Row (1) and (2) are baseline results using ML and MPE training. In this case MPE did not outperform ML because the reason mentioned above. Rows (3) and (4) are results with state merging for all English states merged, or 100% merged, corresponding to the last points in Fig. 3. We see that for English words ML outperformed the baseline results by roughly 5.5% absolute, while MPE did with roughly 7.3% absolute. But it is important that MPE now outperformed ML with state merging by roughly 1.8% absolute. indicating the ability of reducing ambiguity by unit merging. Rows (5) and (6) are results with Gaussian merging, the trend is very similar to rows (3) and (4). But Gaussian level merging was actually significantly better than state level merging. For English words the difference was roughly 1.6% absolute. This indicates that Gaussian is a very good acoustic unit for merging for the problem here, capable of enabling both data-sharing and ambiguity-reduction at the same time.

	Mandarin	English	Overall
(1) ML (Baseline)	83.48	59.90	81.63
(2) MPE (Baseline)	83.19	57.98	81.21
(3) ML (State)	84.25	65.41	82.78
(4) MPE (State)	84.22	67.29	82.89
(5) ML (Gaussian)	84.05	66.67	82.69
(6) MPE (Gaussian)	84.15	68.23	82.90

Table 2. *Overall Experimental Results (Acc.)*

6 CONCLUSION

The distinct nature of code-switched speech is an important issue in the globalized world today, and code-switched speech is actually very frequently observed in the daily lives of many people. In this paper, acoustic unit merging was proposed to help MPE training for transcribing code-switched speech. This approach considers the issues of both data sufficiency and ambiguity between models across different languages. Experimental results showed that the proposed approach actually significantly improved the recognition accuracy, especially for the guest language of English, whose percentage of occurrence was much lower than the host language of Mandarin.

7. REFERENCE

[1] Tanja Schultz and Alex Waibel, "Language Independent and Languzage Adaptive Acoustic Modeling for Speech Recognition", *Speech Communication*, 2001.

[2] Hui Lin, Li Deng, Jasha Droppo, Dong Yu, and Alex Acero, "Learning Methods in Multilingual Speech Recognition", *NIPS*, 2008.

[3] Ching-Feng Yeh, Chao-Yu Huang and Lin-Shan Lee, "Bilingual Acoustic Model Adaptation by Unit Merging on Different Levels and Cross-level Integration", *Interspeech*, 2011.

[4] Ching-Feng Yeh, Liang-Che Sun, Chao-Yu Huang and Lin-Shan Lee, "Bilingual Acoustic Modeling with State Mapping and Three-stage Adaptation for Tran-scribing Unbalanced Code-mixed Lectures", *ICASSP*, 2011.

[5] Ching-Feng Yeh, Chao-Yu Huang, Liang-Che Sun, and LinShan Lee, "An Integrated Framework for Transcribing Mandarin-English Code-mixed Lectures with Improved Acoustic and Language Modeling", *ISCSLP*, 2010.

[6] Yanmin Qian, Daniel Povey and Jia Lu, "State-Level Data Borrowing for Low-Resource Speech Recognition Based on Subspace GMMs", *Interspeech*, 2011

[7] Houwei Cao, Tan Lee and P.C. Ching, "Cross-lingual Speaker Adaptation via Gaussian Component Mapping", *Interspeech*, 2010

[8] David Imseng, Herve Bourlard, Mathew Magimai.-Doss, John Dines, "Language Dependent Universal Phoneme Posterior Estimation for Mixed Language Speech Recognition", *ICASSP*, 2011.

[9] Ching-Feng Yeh, Aaron Heidel, Hong-Yi Lee, Lin-Shan Lee, "Recognition of Highly Imbalanced Code-mixed Bilingual Speech with Frame-level Language Detection based on Blurred Posteriorgram", *ICASSP*, 2012

[10] Daniel Povey, "Discriminative Training for Large Vocabulary Speech Recognition", *PhD thesis, Cambridge University Engineering Dept*, 2003

[11] Ran Xu, Qingqing Zhang, Jielin Pan, Yonghong Yan, "Investigations to Minimum Phone Error Training in Bilingual Speech Recognition", *FSKD*, 2009

[12] Ngoc Thang Vu, Dau-Cheng Lyu, Jochen Weiner, Dominic Telaar, Tim Schlippe, Fabian Blaicher, Eng-Siong Chng, Tanja Schultz, Haizhou Li, "A First Speech Recognition System For Mandarin-English Code-Switch Conversational Speech", *ICASSP*, 2012