Bilingual Acoustic Model Adaptation by Unit Merging on Different Levels and Cross-Level Integration

Ching-Feng Yeh[1], *Chao-Yu Huang*[2], *Lin-Shan Lee*[1,2]

[1] Graduate Institute of Communication Engineering,
[2] Graduate Institute of Computer Science and Information Engineering,
National Taiwan University, Taiwan

andrew.yeh.1987@gmail.com, r98922053@ntu.edu.tw

Abstract

This paper presents a bilingual acoustic model adaptation approach for transcribing Mandarin-English code-mixed lectures with highly unbalanced language distribution. This includes a adaptation structure, merging of Mandarin and English acoustic units on model, state and Gaussian levels, and a cross-level integration scheme. The corpora tested include two real courses, in which special terminologies were produced in the guest language of English (about 15-19%) and embedded in the utterances produced in the host language (about 81-85%). The code-mixing nature of the target corpora and the very small percentage of the English data made the task difficult. Preliminary experiments showed that unit merging was helpful, merging on lower levels offered more improvements, and cross-level integration was even better. The code-mixing situation considered is actually very natural in the spoken language of the daily lives of many people in the globalized world today.
Index Terms: model mapping, bilingual, code-mixing, lecture, acoustic modeling

1. Introduction

In a globalized world today, many people naturally speak more than one language simultaneously. There actually exist two different types of bilingual speech. Code-switching, refers to the case that the speaker switches languages from sentence to sentence. For example, "今天天氣很好. Let's take a walk." (The weather is fine today. Let's take a walk.). In such cases, a language classifier before speech recognition system is very often used and is a good approach. Code-mixing, refers to the case that the speaker mixes terms from different languages in a single utterance. For example, "這是一個 linear algebra 的問題." (This is a problem in linear algebra). In this utterance most general terms are spoken in the speaker's native language (Mandarin Chinese here, referred to as host language), while special terms are spoken in a second language (English here, referred to as guest language). This happens naturally because many special terms do not have translations in the host language, but these terms are of even higher importance for understanding since they are usually key terms of the utterance. In such cases, a language classifier may not be easily applied directly. Because the words in different languages are within the same utterance and produced by the same speaker, it is reasonable to try to build a bilingual acoustic model for the speaker instead.

The simplest solution for such a bilingual acoustic models is to include all acoustic units of the two languages in the phone set. However, very often some acoustic units of the two different languages are very similar, and this may cause ambiguity in recognition since it is difficult to decide to which language a part of the signal belongs. In addition, it is usually relatively hard to collect enough training data for the guest language due to the unbalanced distributions of the two languages, so the acoustic models for the guest language cannot be well estimated. It has been shown that the use of language independent and dependent technologies to take the characteristics of different languages into consideration is important [1], [2], and building a compact bilingual acoustic model for a target speaker is very helpful for many applications such as transcribing lectures.

Many approaches have been proposed to merge multilingual acoustic models by unit on different levels [3], [4], [5], [6], [7], [8]. Comparison of such mapping on different levels for a cross-lingual adaptation task for isolated word recognition was also reported [7].In this paper, unit mapping on different levels (model, state, and Gaussian levels) for bilingual model adaptation for code-mixed lectures is carefully evaluated and analyzed and a cross-level integration approach is proposed. The corpus investigated for transcription is the recorded lectures from two courses offered in National Taiwan University. Although Mandarin was primarily used for these lectures, many utterances were code-mixed with about 81-85% of the signals in Mandarin but only 15-19% in English. The highly unbalanced language distribution makes the task difficult.

2. Adaptation Structure

The overall adaptation structure proposed in this paper is presented here.

2.1. Cascaded Adaptation

The adaptation technique used here is a cascade of three components: a global MLLR [10], a data-driven class-based MLLR and an MAP [9] adaptation. This cascade is referred to as "adaptation" below and will be repeatedly use for three times.

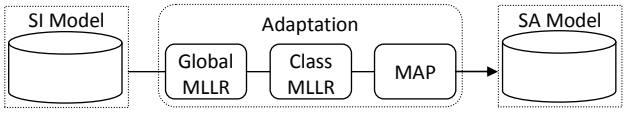

Figure 1: *Cascaded Adaptation.*

2.2. Overall Adaptation Structure

The overall block diagram of the proposed adaptation structure is in Fig.2. We begin with a set of speaker-independent (SI) state-tied triphone models for the complete Mandarin plus English phoneme set respectively trained with Mandarin and English data collected from multiple speakers, referred to as "SI Models (Full)" as shown in the upper left corner of Fig.2. After the cascaded adaptation ("Adaptation (I)" in Fig.2), calculation of distance on different levels between models, states or Gaussians for Mandarin and English are explained below then produces the mapping table at the middle left. This table tells which unit (model, state or Gaussian) is to merge (or share data and parameters) with which corresponding Mandarin unit. Note that English models here are in serious lack of data, so a best mapping unit (model, state or Gaussian) is obtained here for every English unit, which gives a many-to-one

mapping relationship from English to Mandarin. The rest of Fig.2 can be divided into two stages, merging and recovery.

2.2.1. *Merging Stage*

This is the upper right part of Fig.2. Practically merging is always performed on Gaussians by combining the means and covariance matrices on any level. For merging on state level, every Gaussian in the state merges with another Gaussian in the corresponding state with minimum distance. For merging on model level, all states belonging to the model are merged with its corresponding counterpart. Such merging can be performed on a given percentage of English units, producing a set of "shared units", as shown in the upper right corner of Fig.2 as "SI Models (Merged)". The set of models is then retrained with a MAP adaptation using SI training data to reach a steady distribution among SI data. Another cascaded adaptation is then performed, shown as "Adaptation (II)" in Fig.2, in which the merged units are estimated by the data for both languages. In this way, the problem of insufficient data for adapting English model units is properly handled to a good degree by sharing data and parameters with similar Mandarin model units. This gives the speaker adapted (SA) merged model in the lower right corner of this part of Fig.2 as "SA Models (Merged)".

2.2.2. *Recovery Stage*

This is the lower right part of Fig.2. The model merging reduces the impact of insufficient adaptation data, but at the same time limits the achievable likelihood [3]. The merged model units tend to be closer to those of the host language rather than the guest language, because the former dominates the data. This also limits the likelihood between English part of utterances and the merged models. The solution proposed here is to first recover the merged model for both languages and then applying another cascaded adaptation or Adaptation (III) in Fig.2. In the last adaptation process, parameters of models of both languages can be estimated toward their own maximum likelihood. Here the influence of lack of data is much less since the initial parameters for Adaptation (III) have been previously estimated by the common data when they are merged in the merging stage, and are thus more accurate for the target speaker characteristics.

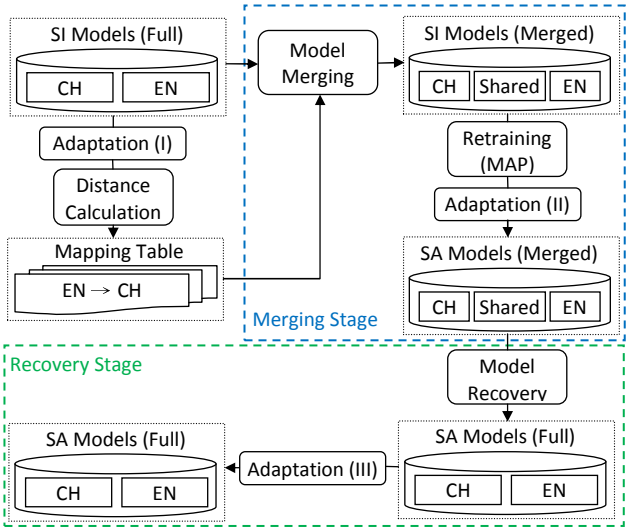

Figure 2: *Adaptation Structure for Proposed Approach.*

3. Distance Evaluation on Different Levels

Data-driven approach with knowledge-based constraints has been proven a good solution to build bilingual models [3], [4], [5], [6], [7], [8] and is used here. Phonemes in the whole Mandarin-English phoneme set are divided into five classes: plosives, fricatives, voiced consonants, vowels and others. Merging is data-driven (between units of minimum distances) but allowed only within the same class.

3.1. Model Level Distance

Phoneme is the minimum unit of sound in a language perceivable by human, while triphone is the best model for phonemes trainable by machines. This is why triphone model merging makes sense. Here a model-based distance between two triphone models is defined. First, for each triphone model, an expected state duration $Dur(S_i)$ in frames is estimated for each state S_i directly using transition probabilities without observation sequences,

$$Dur(S_i) = \sum_{n=1}^{\infty} n[(a_{i,i})^{n-1}a_{i,i+1}] = \frac{a_{i,i+1}}{(1-a_{i,i})^2}, \quad (1)$$

where $a_{i,j}$ is the transition probability from state S_i to state S_j and $[(a_{i,i})^{n-1}a_{i,i+1}]$ is the probability that state S_i last for n frames. This expected duration is then further normalized into $Dur_n(S_i)$ such that the total duration for each triphone model is always 1.0,

$$Dur_n(S_i) = \frac{Dur(S_i)}{\sum_j Dur(S_j)}, \quad (2)$$

where the denominator is the summation over all states in the triphone model. This normalized duration can then be used in aligning two triphone models as in Fig.3.

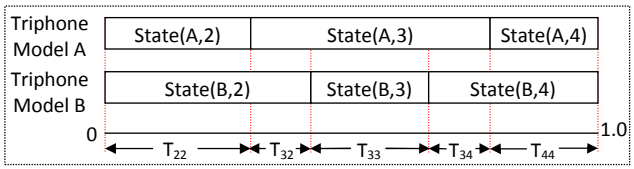

Figure 3: *Alignment of Two Triphone Models.*

Fig. 3 is an example demonstrating the alignment and distance evaluation for two triphone models A and B, each with three states 2, 3 and 4 (1 and 5 are entry and exit states). Note that it is impossible to align observation sequences as in [5] here due to the very limited quantity of adaptation data. In Fig.2, both triphone A and B have a normalized duration of 1.0, and T_{ij} represents the overlapping portion or percentage of state i of triphone A and state j of triphone B, and is used as the weight for the distance between the corresponding states. When evaluating the distance between two states, every state is modeled by a single Gaussian, and the distance between two states is defined as the symmetrical KL divergence between the two Gaussians [3]. So the formula used for estimating the distance between two triphone models is,

$$D_M(m_A, m_B) = \sum_i \sum_j T_{ij} D_{KL}(S_i, S_j), \quad (3)$$

where $D_{KL}(S_i, S_j)$ is defined as in [3], with state S_i, S_j each represented by a single Gaussian. In this way, for every English triphone model, a best mapping candidate of Mandarin triphone model with minimum distance can be found for merging model.

3.2. State Level Distance

There can be limitations in merging triphone models. English and Mandarin are quite different in nature with quite different phoneme sets. So forced merging of distinct triphone models cannot be very smooth. On the other hand, states in HMM are sequential components of phonemes, each with statistically steady distribution for acoustically similar feature vectors, considered corresponding to a certain stage of vocal tract activities. States cannot be perceived by human, but can be well

identified by machine. Although speech production can be very different across languages, it is naturally limited by the physical structure of human vocal tract, which may be well represented by HMM states. This is why we try to merge states here, in which distance between two states is simply the symmetrical KL divergence [3], with each state modeled by a single Gaussian,

$$D_S(S_i, S_j) = D_{KL}(G_i, G_j) \quad (4)$$

where G_i is the single Gaussian mixture that modeling state S_i. In this way, for every state in English triphone models, a best mapped state in Mandarin triphone models with minimum distance can be found.

3.3. Gaussian Level Mapping

Since Gaussian mixtures represent fine structures of state in GMM, merging between Gaussians is certainly possible [7]. The distance between two Gaussians is simple, the symmetrical KL divergence in [3] can be used directly,

$$D_G(G_i, G_j) = D_{KL}(G_i, G_j). \quad (5)$$

However, note that in this way very similar Gaussians simply indicate similar statistical distribution totally in the feature space which is jointly modeled by many Gaussians. Also, here the weights of Gaussian mixtures are ignored, but the weights actually indicate the significance of each Gaussian mixture. W therefore define a distance between Gaussian mixtures constrained by their normalized weight difference,

$$D'_G(G_i, G_j) = \left(b + \left|\frac{w_i - w_j}{w_i + w_j}\right|\right)^{\alpha} D_{KL}(G_i, G_j), \quad (6)$$

where w_i is the weight for Gaussian mixture G_i in the corresponding state, b is a bias parameter to give non-zero distance to dissimilar Gaussians with equal weights, and α is a weight parameter. In this way, Gaussian mixtures tend to be mapped to similar Gaussians with similar weights in corresponding states, or the distributions of the whole feature space for the state better considered.

3.4. Cross-Level Integration

As mentioned above, there exist different advantages and limitations for merging units on different levels, and some extra information may be obtained from distances on different levels. For example, two Gaussians similar on both Gaussian and state level may imply they are more probably produced by similar speech production activities than those with only high similarity on Gaussian level but low similarity on state level. In other words, merging on lower level is finer with higher chance of overfitting, and upper level information can be introduced to reduce such chance. Therefore, for state level mapping, phone-level distance is helpful,

$$D_{CL}(S_i, S_j) = \alpha D_{MP}(P_i, P_j) + \beta D_S(S_i, S_j), \quad (7)$$

$$\alpha + \beta = 1, \quad (8)$$

where $D_{MP}(P_i, P_j)$ is the distance between two monophone P_i, P_j similarity obtained with monophone models. This cross-level distance $D_{CL}(S_i, S_j)$ is probably better than the distance evaluated on state level alone. Similarly, for Gaussian level mapping the cross-level distance below may be better,

$$D_{CL}(G_i, G_j) = \alpha D_{MP}(P_i, P_j) + \beta D_S(S_i, S_j) + \gamma D_G(G_i, G_j), \quad (9)$$

$$\alpha + \beta + \gamma = 1. \quad (10)$$

4. Experiment

4.1. Experiment Setup

The corpora used for the experiments were the recorded lectures of two courses offered in National Taiwan University in form of spontaneous speech with highly unbalanced Mandarin-English code-mixing characteristics as mentioned above.

Table 1. *Experiment Data Characteristics*

	Lecture 1	Lecture 2
Train (hrs)	9.10	7.82
Adapt (mins)	29.86	31.26
Test (I)/(II) (mins)	132.15/29.62	126.21/30.13
Mandarin/English (%)	84.8/15.2	80.5/19.5

The detailed description of experimental data is listed in Table 1. Test (II) is used in the initial experiment for cross-level integration only. The percentage of English is only 15-19%, or roughly 1.5 hours in training and 5 minutes in adaptation data. Therefore, special technologies for such unbalanced bilingual acoustic model adaptation are needed.

The initial SI models were trained from two different corpora for the two languages. The Mandarin models were trained with the ASTMIC corpus of read speech in Mandarin only, produced by 95 males and 95 females, each reading 200 sentences, with a total length of 24.6 hours. The English models were trained with the Sinica L2 Taiwanese English corpus, which was also a read speech corpus in English only but produced by Taiwanese speakers, 229 males and 256 females, with a total length of 59.7 hours. Note that the test set mentioned above was more spontaneous in lecture form, while the SI models were trained with read speech here.

The lexicon used here included English words, Chinese words and all commonly used Chinese characters. The language models were constructed through random forest approach (RFLMs) as in [3], [4]. The recognition accuracy reported below is evaluated also in the same way as in [3] and [4]. That is, when aligning recognition results with reference transcriptions, overlapping of labels from different languages are forbidden, and insertions, deletions, substitutions are calculated for each language and summed up for overall evaluation. The basic unit for alignment is character for Mandarin and is word for English. Due to the unbalanced language distribution, the overall accuracy can be generally regarded as character accuracy.

4.2. Results for Merging Stage

Recognition accuracies for Mandarin, English and overall for the two courses using different versions of acoustic models of the end of merging stage are listed in row (5) - (10) of Table 2. The English accuracy is emphasized here since the English terms are usually key terms for understanding. Row (1) are the results for the initial SI full model without adaptation, which includes all Mandarin-English phoneme set without merging. English accuracy was very poor here. Row (2) are the results for applying the standard cascaded adaptation without unit merging, serving as the baseline of merging stage. We can see the cascaded adaptation is very useful for both languages. Row (4) is the accuracy for speaker-dependent models trained with all training data listed in Table 1, serving as the upper bound. The next several rows are then results when merging is performed on model (labeled M, rows (5)(6)), state (labeled S, rows (7)(8)) and Gaussian (labeled G, rows (9)(10)) level, respectively, in which in each case the first rows ((5)(7)(9), labeled SI) for the mapping table obtained from the initial SI model, while the second rows ((6)(8)(10), labeled SA) for the table from the adapted model as shown in the middle left of Fig.2. We can see the accuracy (particularly for English) was significantly improved by the merging process. In general, the Gaussian level merging (rows (9)(10)) was better than state level (rows (7)(8)) and in turn better than model level (rows (5)(6)), or the lower (and finer) level merging was better, and

the mapping table from the adapted model (rows (6)(8)(10)) was always better.

Table 2. *Merging and Recovery Results (Char Acc.).*

	Course 1			Course 2		
	Mandarin	English	Overall	Mandarin	English	Overall
(1) SI (Full)	53.91	0.39	49.73	41.28	-1.55	37.96
(2) SA (Full, ADP)	80.29	48.30	77.96	71.12	55.63	69.92
(3) SA (Full, ADP*2)	82.05	51.77	79.82	72.17	57.47	71.03
(4) SD (Full)	85.72	64.72	84.08	78.04	72.63	77.63
Merging Stage						
(5) SA (M, SI)	80.12	48.62	77.84	71.06	55.96	69.89
(6) SA (M, SA)	80.16	49.73	77.91	70.59	57.02	69.54
(7) SA (S, SI)	80.23	49.10	77.95	70.68	56.05	69.55
(8) SA (S, SA)	80.57	52.49	78.51	70.62	57.26	69.59
(9) SA (G, SI)	80.34	48.92	78.02	70.88	57.39	69.83
(10) SA (G, SA)	80.97	53.74	78.93	70.85	58.73	69.91
Recovery Stage						
(11) SA (M, Merge)	80.76	50.78	78.57	70.88	57.39	69.83
(12) SA (M, Recover)	80.84	51.81	78.72	71.10	58.69	70.14
(13) SA (S, Merge)	81.56	54.34	79.53	71.18	58.90	70.23
(14) SA (S , Recover)	81.82	58.06	80.04	71.25	60.66	70.43
(15) SA (G, Merge)	82.23	58.88	80.46	71.20	60.79	70.39
(16) SA (G, Recover)	82.45	60.64	80.78	71.29	62.30	70.60

4.3. Recovery Stage Evaluation

The results of adding the recovery stage in the lower part of Fig.2 are in rows (11)-(16) in Table2. Because mapping tables from speaker-adaptive model (rows (4)(6)(8)) are shown better, only such models are considered below. Rows (11)(12) are extensions of row (6) for merging on the model level with mapping table from SA model. Row (9) are results for directly applying the cascaded adaptation on the models after merging without the recovery process in Fig.2, while row (10) on the models with the recovery process exactly as shown in Fig.2. It is clear that a extra stage of cascaded adaptation helped, and the recovery process offered very significant improvements. Similarly, rows (13)(14) are extensions of row (8) and rows (15)(16) of row (10) for merging on state and Gaussian levels. Similar trends can be observed. As shown in Fig.2, actually two cascaded adaptation process were performed repeatedly, Adaptation (II) and (II), so the baseline to be compared here should be that of row (2) followed by an additional iteration of cascaded adaptation. The results are shown in row (3). Comparing row (3) to (2), an additional iteration of cascaded adaptation did bring extra improvements. Comparing rows (11) – (16) with (3), model unit merging was proven useful, and the improvements increases from model level to Gaussian level, the finer the better. Comparing rows (11)(13)(15) to (12)(14)(16), the benefit offered by the model recovery process in maximizing the likelihood between the adaptation data and the recovered models is also significant. The best English accuracy obtained here, 60.64 % and 62.30% in row (16) represents a relative improvement of 17.1% and 8.4% as compared to row (3), and is significantly better.

4.4. Initial Results for Cross-level Integration in Bilingual Modeling

Due to lack of time, only very preliminary experiments were conducted for the cross-level integration, and the initial results are reported here. No adaptation was performed, only bilingual acoustic modeling here. So in Fig.1 all the three times of application of cascaded adaptation were left out, and the recovery stage was left out too. We only started with the speaker dependent model (corresponding to row (4) in Table 2) while tested on Test (II) data of course 1 in Table 1. The results are in Table 3. Row (1) is for the speaker-dependent full model, serving as the baseline. Rows (2)(3)(4) are for merging on the model (M), state (S) and Gaussian (G) levels respectively, in which again we see the English accuracy was significantly improved with improvements increasing from model level to Gaussian level, or the lower or finer level merging the better. We can see here the unit merging is useful not only in adaptation, but in bilingual modeling as well. Rows (5)(6)(7) are the results obtained from cross-level integration (CL) on the state and Gaussian level respectively (the weights α, β, γ also listed). By comparing rows (5) to (3) and (6) to (4), clearly extra improvements were achieved by integrating upper level information

Table 3. *Cross-level Results (Char Acc.).*

	Lecture 1			Parameters		
	Mandarin	English	Overall	α	β	γ
(1) SD (Full)	85.13	62.43	83.41	--	--	--
(2) Merged (M)	85.45	64.11	83.83	--	--	--
(3) Merged (S)	85.10	66.54	83.69	--	--	--
(4) Merged (G)	85.68	69.72	84.47	--	--	--
(5) Merged (S, CL)	85.33	67.02	83.81	0.2	0.8	--
(6) Merged (G, CL)	85.54	70.47	84.40	0.1	0.2	0.7

5. Conclusions

In this paper, a bilingual acoustic model adaptation approach is proposed for highly unbalanced code-mixed Mandarin-English lecture corpora by merging on model, state and Gaussian levels and cross-level integration. Experimental results showed that the proposed approach actually significantly improved the recognition accuracy, especially for the guest language of English, with very limited quantity of adaptation data. Such code-mixing situation is very natural in the spoken language of many people in the globalized world today. Further investigation in this direction is certainly interesting and worthwhile.

6. References

[1] Tanja Schultz and Alex Waibel, "Language Independent and Language 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, Liang-Che Sun, Chao-Yu Huang and Lin-Shan Lee, "Bilingual Acoustic Modeling with State Mapping and Three-stage Adaptation for Transcribing Unbalanced Code-mixed Lectures", ICASSP, 2011.

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

[5] Yanmin Qian and Jia Liu, "Phone Modeling and Combining Discriminative Training for Mandarin-English Bilingual Speech Recognition", ICASSP, 2010.

[6] B. Mak and E. Barnard, "Phone clustering using Bhattacharyya distance", in Proc. Of ICSLP, vol. 4, pp. 2005-2008, Oct. 1996.

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

[8] Yi-Jian Wu, Simon King and Keiichi Tokuda, "Cross-lingual Speaker Adaptation For HMM-Based Speech Synthesis", ISCSLP 2008.

[9] CH Lee, JL Gauvain, "Speaker adaptation based on MAP estimation of HMM parameters", ICASSP, 1993.

[10] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", Computer Speech and Language, 1995

[11] Peng Xu and Frederick Jelinek, "Random forests in language modeling", EMNLP, 2004.

[12] Anoop Deoras, Frederick Jelinek and Yi Su, "Language Model Adaptation using Random Forests", ICASSP, 2010.