RECOGNITION OF HIGHLY IMBALANCED CODE-MIXED BILINGUAL SPEECH WITH FRAME-LEVEL LANGUAGE DETECTION BASED ON BLURRED POSTERIORGRAM

Ching-Feng Yeh[1], Aaron Heidel[2], Hong-Yi Lee[1], 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

ABSTRACT

In this work, we proposed a new framework for recognition of highly imbalanced code-mixed bilingual speech using an additional frame-level language detector in the conventional recognition system. Blurred posteriorgram features (BPFs) are also proposed to be used in the language detector. The approach was evaluated with real spontaneous lectures offered at National Taiwan University. The highly imbalanced language distribution in code-mixed speech makes the task difficult. Preliminary experimental results showed not only very good performance improvement, but the improvement is complementary to that brought by better acoustic models, whether due to better adaptation approach or increased training data. The code-mixed bilingual speech is frequently used in the daily lives of many people in the globalized world today.

***Index Terms*— code-mixing, multilingual, ASR**

1. INTRODUCTION

In the era of a globalized world today, the numbers of people speaking more than one language grow every day. It is therefore important to construct multilingual speech recognition technologies rather than simply the conventional monolingual technologies. Previous works [1], [2], [3], [4], [5] showed the distinct characteristics of such multilingual speech recognition technologies. In this paper, we proposed a new recognition framework of recognizing highly imbalanced code-mixed bilingual speech.

Bilingual speech can be in general classified into two categories. One is code-switching, in which the speaker switches languages from sentence to sentence. For example, the sentences “It’s fine. 謝謝你. (It’s fine. Thank you)”, where the first sentence is in English, while the second in Chinese. The other is code-mixing, in which the language are switched from words to words. For example, “這個 equation 很複雜. (This equation is very complicated.)”, in which the word “equation” in the guest language of English embedded in a sentence in the host language of Chinese. Such code-mixed speech is very commonly used by people with non-English native languages in their daily lives (with English as the guest language and the native language as the host), especially when many English words are not properly translated into their native languages. The latter category of code-mixed speech is the target of this paper.

In code-mixed speech, the language distribution is usually highly imbalanced, that is, the speaker tends to use primarily the native language as the host language in constructing the sentence, but with only a few words in foreign language embedded as the guest language. As a result, the occurrences of the guest language are much less than those of the host language. A direct impact of such phenomenon is that the recognition accuracy for words in guest language is usually much lower because of the relatively poor acoustic and language models for the guest language with much less data. This is a serious problem because the words in guest language are usually special terminologies, new words or key terms, and therefore important.

It is well known that language identification is very helpful in recognizing multilingual code-switched speech (languages are switched from sentence to sentence), and many successful approached have been proposed. But these approaches cannot be directly applied in the code-mixed problem considered here. First, the basic unit for language identification for code-switched speech is usually the sentence, since languages are switched between sentences. But in code-mixed speech considered here, the languages are switched within the sentence, so the basic unit for language identification should be word, sub-word units or even a frame. Besides, in code-mixed speech, the two languages are usually produced by the same speaker, and the words in guest language are usually pronounced in the style of host language, i.e., with phonemes and prosody almost the same as the host language. This makes the acoustic signals for host and guest languages very similar and difficult to distinguish.

In this paper, we propose to use a special designed frame-level guest language detector using blurred posteriorgram features in a new recognition framework to identify the guest language frames, although acoustic models can also be improved by merging and recovery algorithms previously [3], [4]. Significant improvements were achieved in the experiments.

2. LANGUAGE BIAS PROBLEM OF CODE-MIXED SPEECH

2.1. Testing Environment

The corpus used for the experiments reported here were the recorded spontaneous Chinese-English lectures of two

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

courses (Course 1 and 2) offered in National Taiwan University, with Chinese as the host language and English as the guest language. This corpus was divided into training, adaptation and testing sets as listed in Table 1, where we can see the percentage for English is only 15-20%. So the language distribution is highly imbalanced.

Different sets of acoustic models were used here. In the case that very limited adaptation data are available, we used the standard techniques for model adaptation [3] started with a set of speaker-independent (SI) models, which give the first set of speaker-adapted (SA-1) models. When the recently proposed special approaches for adapting code-mixed bilingual acoustic models [3] can be used by merging similar model units from the two languages, an improved set of SA models (SA-2) can be obtained. When the large training set in Table 1 is available, a speaker-dependent (SD) model can also be trained. These three sets of acoustic models, SA-1, SA-2 and SD, will be used in all experiments reported below.

The SI models were trained by two separate corpora, ASTMIC and TWNAESOP, both recorded with multiple speakers and are gender-balanced with a total length of 74.3 hours [3].

Table 1. *Details for the Target Corpus*

	Course 1	Course 2
Training Set (hrs)	9.10	7.82
Adaptation Set (mins)	29.86	31.26
Testing Set (mins)	132.15	126.21
Mandarin/English (%)	84.8/15.2	80.5/19.5

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. Chinese words were also generated by segmenting a large corpus using PAT-Tree base approaches. We used the Kneser-Ney tri-gram model started with a background model and then adapted with training set for the target lecture here.

2.2. Imbalanced Frame-level Language Identification with a Conventional Recognizer

A bilingual speech recognizer can be easily constructed by expanding the conventional monolingual recognizer using bilingual acoustic / language models and bilingual lexicon to transcribe code-mixed speech. The frame level language identification accuracies for initial experiments for such a conventional bilingual recognizer are listed in Table 2. The three sets of acoustic models mentioned above (SA-1, SA-2 and SD) are tested on the two courses in Table 1. Here the precision and recall are obtained by comparing with the forced alignment results of manual transcription. From Table 2, it is clear that the precision and recall for English are much lower than those for Chinese regardless of the type of acoustic models, especially the values of recall. For example, with SA-2 model, only 68% of English frames were recognized as English words while the other 32% were recognized as Chinese words. This is the language bias problem in the code-mixed environment, i.e., the system tends to take every speech segment as a part of Chinese word. So, many English words are recognized as Chinese words, naturally due to the highly imbalanced language distribution. Because of much smaller quantity of available data for the guest language, the acoustic and language model parameters for the guest language cannot be estimated very well. This leads to the proposed framework below.

Table 2. *Frame-level Language Identification Accuracy with a Conventional Recognizer*

Acoustic Models	Course 1				Course 2			
	Mandarin		English		Mandarin		English	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
SA-1	0.93	0.99	0.85	0.53	0.94	0.95	0.63	0.58
SA-2	0.95	0.97	0.77	0.68	0.96	0.92	0.55	0.71
SD	0.96	0.99	0.90	0.72	0.96	0.98	0.81	0.72

3. NEW RECOGNITION FRAMEWORK FOR CODE-MIXED SPEECH

The new recognition framework proposed here to handle the above problem is presented here.

3.1. Overall Picture of the New Framework

The basic idea proposed here is to add a frame-level guest language detector to the conventional recognition framework as shown in Fig. 1, and then boost the recognition scores for those frames identified as in guest language for guest language phoneme models. For the corpus tested here, the goal of this guest language detector is to detect English frames that the conventional recognizer may not be able to identify. Everything else in Fig.1 is exactly the same as in a conventional recognizer, except everything is bilingual. Assume the guest language detector generates a posterior probability of guest language given each feature vector o_t for a speech frame at time t , $P(G|o_t)$, and a probability of host language given o_t , $P(H|o_t)$, where $P(G|o_t) + P(H|o_t) = 1$. The acoustic model score for frame o_t with an HMM state q_j , $P(q_j|o_t)$, can then be boosted into a new score $\hat{P}(q_j|o_t)$ as below,

$$\hat{P}(q_j|o_t) = \begin{cases} P(q_j|o_t) \times \left[\dfrac{P(G|o_t)}{1-P(G|o_t)}\right]^{\alpha} & if P(G|o_t) > 0.5 \\ & and\ q_j \in G \\ P(q_j|o_t) & otherwise \end{cases} \tag{1}$$

where $\hat{P}(q_j|o_t)$ is the score to be used in the recognizer, G is the set of all HMM states for guest language phoneme models, and α is a parameter. If a frame o_t is identified as in guest language, or $P(G|o_t) > 0.5$, its score with states of guest language phoneme models are boosted according to the posterior probability $P(G|o_t)$ from the detector, otherwise the score is not changed. Because the conventional recognizer can already choose host language models very well, no action is needed if $P(H|o_t) > 0.5$.

The above frame-level guest language detector can be implemented in different ways. It has been shown [6] that this can be achieved by neural network with properly chosen features, such as MFCCs with longer context. In this work, we also propose to use the neural network, but with different features. Since bilingual speakers usually tends to pronounce guest language words using host language phonemes, so the MFCC features for the two languages are very similar, not very useful in our detector. We therefore propose blurred posteriorgram features (BPFs) extracted from decoded lattices as presented below.

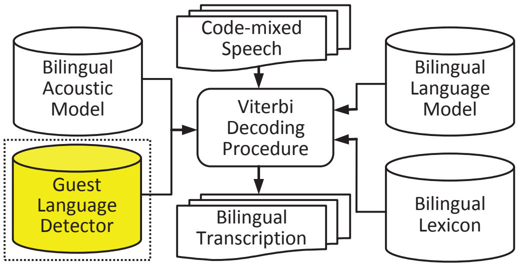

Figure 1. *Proposed Recognition Framework for Code-mixed Bilingual Speech*

3.2. Blurred Posteriorgram Features (BPFs)

Here we wish to find a good set of features based on the posterior probability distribution for each speech frame to be used as the input to the neural network for the guest language detector. Each utterance is first decoded into a phoneme lattice in the first-pass recognition, where the phoneme set includes phonemes for both the host and guest languages. With this phoneme lattice each frame o_t has an N-dimensional posteriorgram vector $P_t = \{P(p_i|o_t), i = 1, 2 \ldots \mathrm{N}\}$, where p_i is a phoneme in either host or guest languages, N is the total number of phonemes for the two languages, and $P(p_i|o_t) = 0$ for those phonemes p_i not appearing in the lattice at time t.

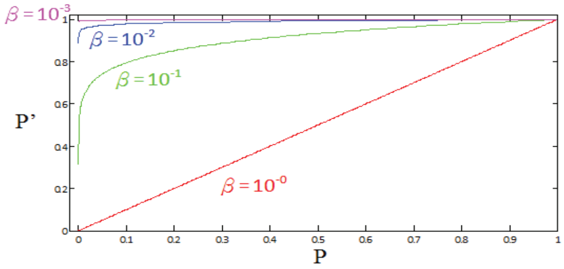

Figure 2. *The Concept of Blurring Transformation*

The problem here is that very often guest language phonemes are decoded as host language phonemes, or $P(p_i|o_t)$ is usually relatively lower for guest language phonemes even if o_t belongs to a guest language phoneme. So we try to "blur" the posteriorgram,

$$P'(p_i|o_t) = P(p_i|o_t)^{\beta}\ , 0 < \beta < 1, \qquad (2)$$

where β is the blurring factor, much smaller than 1 and close to 0. The concept of (2) is shown in Fig.2 for a few selected values of β, where we see $P'(p_i|o_t)$ is monotonically increasing for increasing $P(p_i|o_t)$ while all posterior probabilities $P(p_i|o_t)$ are moved towards unity in a non-linear manner. In other words, $P(p_i|o_t)$ is significantly increased if it is small (or very possibly p_i belongs to the guest language), while only very slightly increased if $P(p_i|o_t)$ is large (or very possibly p_i belongs to the host language). In this way the posterior probabilities for the guest language is properly enhanced, while the posteriorgram is "blurred". These blurred posteriorgram features (BPFs) are then used as the input to the neural network for the guest language detector with two targets: guest language or not.

4. EXPERIMENTAL RESULTS

The experimental results consist of two parts, one for language detection, and the other for speech recognition.

4.1. Language Detection Experiment

In this experiment, the language identification performance in terms of precision and recall for English part for the guest language detectors using MFCCs and the proposed blurred posteriorgram features (BPFs) were evaluated and compared with the conventional recognizer. The acoustic models SA-1 and SD listed in Table 2 were used. The results are listed in Table 3.

Table 3. *Language Detection Experimental Results*

		Course 1 (English Part)		Course 2 (English Part)	
		Precision	Recall	Precision	Recall
Recog.	(1) SA-1	0.85	0.53	0.63	0.58
	(2) SD	0.90	0.72	0.81	0.72
MFCC	(3) SA-1	0.28	0.50	0.32	0.45
	(4) SD	0.39	0.68	0.44	0.51
BPF	(5) SA-1 (β = 1.0)	0.85	0.49	0.85	0.51
	(6) SA-1 (β = 0.1)	0.83	0.52	0.86	0.55
	(7) SA-1 (β = 0.01)	0.70	0.58	0.81	0.62
	(8) SA-1 (β = 0.001)	0.69	0.56	0.72	0.53
	(9) SD (β = 0.01)	0.93	0.74	0.82	0.70

In Table 3, rows (1) and (2) are the results using the conventional recognizer serving as the baseline, actually copied from the first and the last rows in Table 2. Rows (3) and (4) are the results using MFCCs with longer context as the features for neural network, as proposed previously [6]. We can see for bilingual speakers, it is difficult to extract language information using MFCCs as features. Although the recalls are close to the baseline (rows (3)(4) vs. (1)(2)), the precisions are very low. Rows (5) ~ (9) are then the results using proposed blurred posteriorgram features (BPFs) with various values of β. According to this result we selected 0.01 as the β value for the following experiment. Here we see improved recalls together with precision either improved or degraded slightly (rows (5) ~ (9) vs. (1)(2)). We also wish to find out the ability of the guest language detector in detecting guest language frames which the conventional recognizer could not identify previously. This can be observed with two parameters, the additional true acceptance rate (Add.TA, percentage of extra correctly detected guest language frames) and additional false alarm rate (Add. FA, percentage of extra incorrectly detected guest language frames) as compared to the conventional recognizer. The results listed for the guest language detector using MFCCs and the proposed

BPFs are respectively listed in Table 4. We see for both MFCCs and BPFs significant additional true acceptance rates for both MFCCs and BPFs. However, the additional false alarm rates were also high for MFCCs but very low for BPFs. The high Add. TA rate and low Add. FA rate for BPFs indicated that BPFs can be good features for the problem considered here.

Table 4. *Add. TA and FA for the Guest Language Detector*

	Acoustic Models	Course 1 (English Part)		Course 2 (English Part)	
		Add. TA	Add. FA	Add. TA	Add. FA
MFCC	(1) SA-1	0.19	0.19	0.13	0.13
	(2) SD	0.09	0.10	0.08	0.09
BPF	(3) SA-1	0.15	0.02	0.11	0.02
	(4) SD	0.07	0.01	0.07	0.02

4.2. Speech Recognition Experiment

In this experiment, we tested the proposed recognition framework which integrated the guest language detector using BPFs as features. The results are listed in Table 5. The way the recognition accuracy was evaluated followed the earlier work [3]. That is, when aligning recognition results with the reference transcriptions, insertions, deletions, substitutions are evaluated respectively for each language and summed up for overall evaluation. The basic unit for alignment is character for Chinese and word for English.

Table 5. *Speech Recognition Accuracy Results*

		Course 1			**Course 2**		
		Mandarin	English	Overall	Mandarin	English	Overall
SA-1	(1) Recog.	74.42	41.08	71.81	69.44	53.37	68.20
	(2) BPF	75.55	47.54	73.35	70.02	59.82	69.24
	(3) UB	76.87	56.57	75.28	71.19	67.95	70.94
SA-2	(4) Recog.	77.60	48.24	75.30	71.25	58.82	70.29
	(5) BPF	77.67	50.51	75.54	70.85	60.33	70.04
	(6) UB	78.78	56.85	77.06	71.87	64.89	71.33
SD	(7) Recog.	83.80	62.40	82.13	77.40	72.18	77.00
	(8) BPF	84.16	65.65	82.71	77.54	73.65	77.24
	(9) UB	84.88	71.95	83.87	78.15	78.30	78.16

In Table 5, rows (1), (2) and (3) are the results using the standard adapted acoustic models (SA-1) mentioned in section 2. Row (1) is the results for the conventional recognizer (Recog.) simply using bilingual acoustic / language models and a bilingual lexicon. Row (2) is the results for the proposed approach as the framework in Fig.1 including the guest language detector using the proposed blurred posteriorgram features (BPFs). We can see the English accuracy was improved significantly, while the Chinese accuracy was improved too. Row (3) is the oracle results with perfect guest language detection obtained with forced alignment with the reference transcriptions, serving as the upper bound (UB). We can see there is still quite good space for further improvement. Rows (4)(5)(6) are exactly the same as rows (1)(2)(3), except with acoustic models adapted using the recently proposed special approaches for highly imbalanced code-mixed speech (SA-2). We can see the trends are the same as in rows (1)(2)(3). This also shows the approaches proposed here are equally useful for different acoustic models, and the improvements brought by improved model adaptation and by the approaches proposed here are additive and complementary to each other. Rows (7)(8)(9) are again exactly the same as rows (1)(2)(3), except with the speaker-dependent model (SD), offering exactly the same observations. So the proposed approaches are complementary to not only improved model adaptation, but increased training data as well. Better performance was always achievable. We also noticed that the improvement is more significant for weaker acoustic models, which is reasonable. But we did not tune the parameters deliberately. In all these experiments, α in (1) was fixed to 1.0 and β in (2) to 0.01, although the best values for these parameters were not known at all.

5. CONCLUSION

In this work, we proposed a new recognition framework for highly imbalanced code-mixed speech and a method for guest language detection using blurred posteriorgram features (BPFs). The experiments were performed over real lectures given at National Taiwan University. The distinct nature of highly imbalanced language distribution in code-mixed speech makes the task difficult. The performance improvement offered by the proposed approach was very good, and the potential of the proposed framework was also analyzed. The code-mixed speech investigated in this work is very frequently observed in the daily lives of many people in the globalized world today.

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, 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 Transcribing 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] David Imseng, Herve Bourlard, Mathew Magimai.-Doss, John Dines, “Language Dependent Universal Phoneme Posterior Estimation for Mixed Language Speech Recognition”, ICASSP, 2011.

[7] 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

[8] CH Lee, JL Gauvain, “Speaker adaptation based on MAP estimation of HMM parameters”, ICASSP, 1993.