Code-switching Speech Detection Method by Combination of Language and Acoustic Information

Hongji Zhang
School of Computer Science and Technology, Beihang University
Beijing, 100191,China
E-mail: zhang_hong_ji@163.com

***Abstract*—In this paper, we propose a new speech detection method to English-Mandarin code-switching speech. Unlike previous methods, in this method we first train a support vector machine (SVM) model based on feature parameters and Gaussian Mixture Model (GMM) , then integrate the language identification (LID) information based on SVM model and acoustic information into the decoding process. Lastly, we develop a prototype system to present the method. Experiments proved that our method we can improve the accurancy of code-switching speech recognition at a certain degree compared with previous methods.**

Keywords-code-switching speech; acoustic model; language identification; support vector machine

I. INTRODUCTION

Code-switching is defined as the use of words from two different languages within a single discourse or within a single utterance. It is a common phenomenon in many bilingual societies. [1][2] In Hong Kong code-switching between Cantonese and English is used widely in many occasions. In Singapore Mandarin and English are frequently mixed and spoken in daily conversations. Recognizing an utterance in mixed languages has still been a challenge for the present automatic speech recognition (ASR) systems. The main difference between ASR of code-witching utterances and monolingual or multilingual utterances is that there are multiple languages within a single utterance.

Our goal in this research is to develop a method that could recognize both Mandarin and English in a code-switching speech and give the switching times when the speech is over. In this paper, we propose the methods of model training, including the support vector machine(SVM) model training using in language identify(LID). Unlike the previous studies, we integrate information from language identification into the decoding process, using the combination of the acoustic model (AM) score and the LID score to identify the phoneme. And the experiments proves that we can achieve the correctly ration. By the way, we will use the CMU sphinx toolkit in our method.

The rest of this paper is organized as follows: Related works are introduced in Section 2. Model, the new proposed method and system are described in Section 3. In Section 4 the experiments and achieved results are presented. Finally, we draw some conclusions in Section 5.

II. RELATED WORK

Automatic speech recognition in Mandarin and English has already been studied by many researchers [3][4], but code-switching between these two languages is not much applied in ASR [5]. There have been two different approaches to code-switching recognition. The first way is involving a language boundary detection (LBD) algorithm that divides the input utterance into language-homogeneous segments. In [6], the author detects the boundary from English and Cantonese code-switching utterances by using bi-phone probabilities, which were calculated to measure the confidence that the recognized phones are in Cantonese. Another paper reported that the use of LSA-based Gaussian Mixture Model (GMM), VQ-based bi-gram language model and a likelihood ratio hypothesis test could be efficient to determine the optimal number of language boundaries. [7]

The second way aims to develop a cross-lingual speech recognition system, which can handle multiple languages in a single utterance. The acoustic models, language models, and pronunciation dictionary are designed to be multi-lingual and cover all languages concerned.

There are also some other papers on code-switching detection. In [8], automatic recognition of Mandarin-Taiwanese code-switching speech was investigated. It was found that Mandarin and Taiwanese share a large percentage of lexicon items. Their grammar was also assumed to be similar. A one-pass recognition algorithm was developed using a character-based search net. A mixed-lingual keyword spotting system was developed in [9] for auto-attendant applications. The keywords to be detected could be in either English or Chinese.

III. DETECTION METHOD BY THE COMBINATION OF LANGUAGE AND ACOUSTIC INFORMATION

In the following sections, the phrase “code-switching speech” means English-Mandarin code-switching speech.

A. Acoustic Model

Because there are more than one language phonemes in code-switching speech, we first need to establish a comprehensive, suitable phonemes table and then make phoneme models. Both knowledge-based and data-driven are the methods of cluster the English phonemes and Mandarin phonemes together. Here we use the knowledge-based

Published by Atlantis Press, Paris, France.

method. We count the phonemes in English and Mandarin respectively and merge them based IPA .Now we get a hybrid phoneme table including 60 phonemes, in which all the phonemes are vowels and consonants.

We then establish the acoustic models. In this paper three acoustic models are established. Language-dependent models include both Mandarin acoustic model and English acoustic model。Language-independent model, i.e. mixed multilingual acoustic model, is established using the hybrid phoneme table above. Considering the coarticulation in continuous speech, we use triphone to establish the model. That is, considering a phoneme by itself, its left and right adjacent phonemes.

For English, we can directly use the acoustic model provided by CMU Sphinx. For Mandarin and mixed language, we use CMU Sphinx SphinxTrain tool to train the acoustic model.

B. Language Model

Language models are established by N-gram rule. In this paper, N is2 or 3, that is, we use Bi-Gram and Tri-Gram models. The probability that current word appears is determined by one or two words before it. The probability is defined as follows:

P(w1) probability of word w1 appears

P(w2| w1) probability of word w2 appears when w1 appeared before

P(w3| w2, w1) probability of word w3 appears when w1, w2 appeared before

For English, we can directly use the language model provided by CMU Sphinx. For Mandarin and mixed language, we use CMU Sphinx Cmuclmtk [10] tool to train the language model.

C. Dictionary

Similarly, the code-switching dictionary file also contains multilingual words, and each word has a phoneme sequence correspondingly. Now we use the English dictionary file and Mandarin dictionary file provided by CMU Sphinx and merge them into one file as a code-switching dictionary file.

D. Support Vector Machine Model

Language identification is a binary classification problem, so support vector machine (SVM) is suitable to solve this task. In this subsection we briefly describe SVM on mathematical, and then point out how to use SVM in this paper.

The linear discriminative function in d-dimensional space is

$$g(x) = w \bullet x + b \tag{1}$$

With the normalized discriminative function, the classifier interval is converted to $2/\|w\|$, The problem of solving the optimal hyperplane is transformed into

$$\min_{w,b} \frac{1}{2} \| w \|^2 \quad s.t. \; y_i \left[w \bullet x_i + b \right] \geq 1 \tag{2}$$

According to the Lagrange discriminative function, the optimal classification hyperplane is

$$f(x) = \mathrm{sgn}\left\{ \left(w^* \bullet x + b^* \right) \right\} = \mathrm{sgn}\left\{ \sum_{i=1}^{n} \alpha^* y_i \left(x_i \bullet x \right) + b^* \right\} \tag{3}$$

where α^* is the support vector, b^* is the judgment threshold.

The key of the optimal classification machine is the selection or construction of kernel function, there are four kinds of kernel functions for SVM, linear function, polynomial function, RBF function and Sigmoid function. In this paper, we choose the RBF kernel function.

The training sample of language identification is very large. If we use the feature parameters to train SVM model directly, it is not only unrealistic, but also difficult to guarantee the convergence of the model. Therefore, we have adopted the GMM as SVM training objects, which can effectively reduce the number of training samples, but also keep robustness.

We process the MFCC parameter coming from feature extraction step and get an abstract GMM model which can describe the language feature distribution. Because the mean parameter of GMM model representatives the relative position of the different languages feature in the feature space, then we choose the mean parameter as the training data of SVM language classification model.

We use 1-to-1 training strategy, the target language and the impersonating language are English and Mandarin .We establish two SVM models for them respectively.

E. Detection Method

Acoustic model belongs to the generative model while SVM model is a discriminative model, there is good complementarity between the two models, and the fusion of the two models should be able to improve the speech recognition performance. In this paper, we propose a new method of information combination. The steps to do speech recognition on this task could be described in the following: Firstly, we will do feature extraction to the wave file/voice and get a feature parameter sequence. Secondly, a voice activity detector is used to separate speech and non-speech segments in each utterance. The speech segments are then evaluated by acoustic model and LID classifiers to produce two log likelihood scores for each speech frame. Thirdly the post processing produces a linear combination of the acoustic model score and the LID score with different weight value. The decoding then proceeds using the combination score instead .Lastly, we use language model to determine the recognition result and give the code-switching times.

F. Prototype System

We develop a system to implement code-switching speech recognition using the method we have mentioned in E subsection. The system consists of model training module and recognition module.

Model training module complete front-end training and provide the models and tools, Including acoustic models, language models and SVM model for languages classification. Recognition module mainly includes feature

extraction, voice activity detection, language classification and a post processing procedure.

The system processes as follows:

Training module

Step1: Doing feature extraction to the training speech, and getting a feature parameter sequence;

Step2: Processing the feature parameter to obtain the GMM model by EM algorithm, labeling the mean vector parameter of English GMM model +1 and labeling the mean vector parameter of Mandarin GMM model -1;

Step3: Training the GMM model, and getting a SVM model of language identification;

Step4: Training the acoustic model and language model of Mandarin speech and code-switching speech. Building the dictionary file.

The training process of SVM model is shown in Figure1.

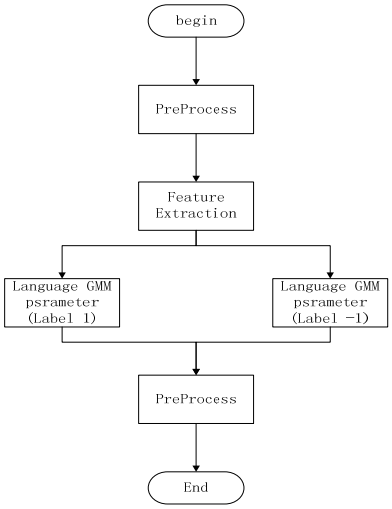

Figure 1. SVM model`s training process

Recognition module

Step1: Doing feature extraction to the input speech, and getting a feature parameter sequence X and GMM model;

Step2: Matching the GMM mean vector parameter with SVM model to obtain the language likelihood score:

$$Scor_{LID} = p(\lambda_i \mid X) = \frac{p(X \mid \lambda_i)\, p(\lambda_i)}{p(X)} \tag{4}$$

Where $\lambda_i \ (i = 1, 2)$ is the GMM model parameter of English or Mandarin;

Step3: Doing voice activity detection to separate speech and non-speech segments in each utterance;

Step4: Recognizing the phone in each speech frame and getting the acoustic likelihood score.

$$Scor_{AM} = p(Pho_i \mid X) \tag{5}$$

Step5: First step in post processing, combining the language likelihood score and acoustic likelihood score as below:

$$Scor = \omega * Scor_{LID} + (1-\omega) * Scor_{AM} \tag{6}$$

Where ω is equal or less than 0.25.

Take k highest scores as the candidate results

Step6: Second step in post process, identify the candidate results by language model and take the candidate result with highest score calculated by step5 as final recognition result if its language model scores is higher than the threshold $Thre_{LM}$. Otherwise, turn to step 7.

Step7: Take the candidate result with highest score as final recognition result.

Step8: Give the speech recognition result and code-switching times.

The structure of proto system is shown in Figure2.

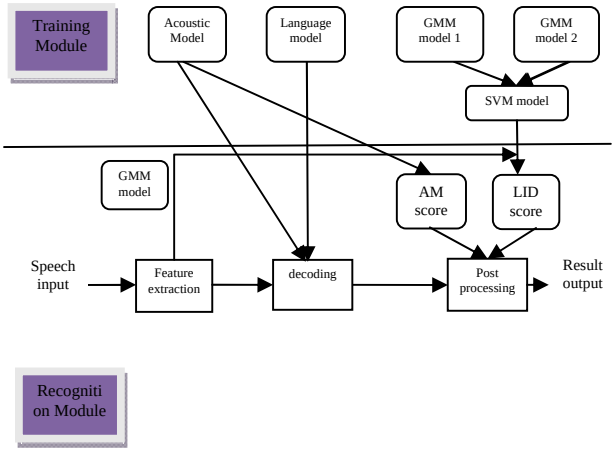

Figure 2. Structure of prototype system

IV. EXPERIMENT AND RESULT EVALUATION

A. *Corpus*

Mandarin corpus: This paper uses the CASIA98-99 speech corpus which is often used in the study of the universal, large vocabulary, continuous speech recognition engine. The CASIA 98-99 corpus consists of the voices of 10 females and 33 males from the 863 voice library. Every subject read 60 sentences and each sentence contains 5-15 words. All the contents are Chinese and are stored in Chinese characters. The paper uses 1680 sentences to train the model.

English corpus: The English corpus that we use in this paper is the Resource Management (RM1) database recommended by CMU Sphinx. RM1 is offered by DARPA to design the voice databases to evaluate the continuous speech recognition systems for researchers. The RM1

database contains 160 individuals' samples and 2500 sentences. The paper uses 1132 sentences to train the model.

Code-switching speech corpus: We use SEAME corpus. SEAME is a conversational Mandarin-English code-switching speech corpus recorded from Singaporean and Malaysian speakers. The corpus is designed for multiple research purposes which include language boundary detection, language identification studies and multilingual Large Vocabulary Continuous Speech Recognition (LVCSR) systems. The average number of code-switches within each utterance is 2.6 when counting only switches between Mandarin and English. The corpus contains 9,210 unique English and 7,471 unique Mandarin words. The duration of monolingual segments is very short: More than 82% English and 73% Mandarin segments are less than 1 second long. We divide the corpus into two sets (training and test set) and distribute the data based on several criteria. Table 1 lists the statistics of the SEAME corpus in these two sets.

All the audios in these three corpuses were sampled at 16 kHz with a resolution of 16 bit, saved as .wav file.

B. Result Evaluation Standard

We choose two evaluation metrics to evaluate the algorithm proposed by this paper, Word Error Rate (WER) and accuracy as follows.

WER: If the detected speech consists of N words and I words are inserted, D words are deleted and S words are replaced in the results, then WER is

$$WER = (I+D+S)/N \quad (7)$$

Similarly, accuracy is defined as

$$Accuracy = (N-D-S) / N \quad (8)$$

Meanwhile, given that the characteristics of language switching, we choose an additional metric, i.e., the accuracy of detecting language switching.

C. Result

Table2 and table 3show the result in different conditions.

As shown in table 2, the system performance improves significantly as acoustic score increases, and reaches the maximum at the point where acoustic score is 0.9 and language score is 0.1. But as acoustic score continues to increase, the performance degrades. So, the choice of weights has a significant effect on the system performance.

Table 3 shows the results of our algorithm and the algorithms in [6], [7]. As shown, our algorithm improves the rate of detecting language switching, for example, accuracy increases by 11.7%, 5.7% and 8.1%, respectively and WER decreases by 12.6%, 6.2% and 8.3%, respectively.

TABLE I. STATISTICS OF THE SEAME CORPUS

	Train Set	**Eval Set**	**Total**
Speakers	139	18	257
Duration(hours)	58.4	4.4	62.8
Utterances	48040	4105	52245

TABLE II. RESULTS OF DIFFERENT PARAMETERS

Experiment	**weight**	**accuracy of language swithing (%)**	**Accuracy (%)**	**WER (%)**
Experiment 1	0.95 0.05	79.5	76.2	26.9
Experiment 2	0.9 0.1	82.1	79.1	22.7
Experiment 3	0.8 0.2	75.3	72.6	30.1
Experiment 4	0.75 0.25	72.9	70.8	32.6

TABLE III. RESULTS OF DIFFERENT METHODS

Detection Method	**accuracy of language swithing (%)**	**Accuracy (%)**	**WER (%)**
Bi-phone probability （knowledge based）	69.62	67.4	35.3
Bi-phone probability (data driven)	76.54	73.4	28.9
GMM-LSI	74	71	31
Combination of LID and AM Information	82.1	79.1	22.7

V. CONCLUSION

In this paper, we propose a new method for spontaneous English-Mandarin code-switching speech and have presented it in a LVCSR system. Some techniques are used to enhance our system. For acoustic modeling, we applied two phone set merging approaches based on IPA. On language model level, we used n-gram rule based on statistical model. We also used SVM method on language identification. Furthermore we integrated the language identification information into the decoding process by using a linear combination of the likely hood scores approach. The experiments approved that our system achieves the accuracy to 82.1% and reduces the WER to 22.7% .

REFERENCES

[1] B. H. Chan, "Code-mixing in Hong Kong Cantonese-English bilinguals: Constraints and process," CUHK Papers in Linguistics, 4, pp. 1-24, 1993.

[2] P. Li, "Spoken Word Recognition of Code-Switched Words by Chinese-English Bilinguals," Journal of Memory and Language. 35, pp. 757 - 774, 1996.

Article in a conference proceedings:

[3] P. C. Ching, et al., "From phonology and acoustic properties to automatic recognition of Cantonese," ISSIPNN- 94, Hong Kong, pp. 127 - 132, 1994.

[4] K.F. Chow, T. Lee and P.C. Ching, "Sub-syllable acoustic modeling for Cantonese speech recognition," Proc. Of ISCSLP 1998, Singapore, pp. 327 - 342, 1998.

[5] P. Fung, X. H. Liu, and C. S. Cheung, "Mixed language Query Disambiguation," Proc. Of ACL 1999, Maryland, pp. 333 - 340, 1999.

[6] J. Y. C. Chan, et al., "Detection of Language Boundary in Code-switching utterances by Bi-phone Probabilities," ISCSLP 2004, Hong Kong, pp. 293-296, 2004.

[7] C. J. Shia, Y. H. Chiu and C. H. Wu, "Language Boundary Detection and Identification of Mixed-Language Speech Based on MAP Estimation," ICASSP, Montreal, Vol.1 pp. 381-384, 2004.

[8] D. C. Lyu, R. Y. Lyu and C. N. Hsu, "Speech recognition on code-switching among the Chinese dialects," ICASSP, Toulouse, Vol.1 pp. 1105-1108,2006.

[9] S. R. You , et al. , "Chinese-English mixed-lingual keyword spotting," ISCSLP 2004, Hong Kong, pp. 237-240, 2004.

[10] http://www.speech.cs.cmu.edu/tools/lmtool.html.