Language Identification on Code-Switching Utterances Using Multiple Cues

Dau-Cheng Lyu [1] *and Ren-Yuan Lyu* [2]

[1] Department of Electrical Engineering, Chang Gung University, Taiwan
[2] Department of Computer Science and Information Engineering, Chang Gung University, Taiwan
d9221003@stmail.cgu.edu.tw, rylyu@mail.cgu.edu.tw

Abstract

Code-switching speech is an utterance containing two or more languages. Usually, the switching linguistic unit is in clause or word levels. In this paper, a two-stage framework is proposed, containing a language identifier and then a speech recognizer, to evaluate on a Mandarin-Taiwanese code-switching utterance. In the language identifier, we use multiple cues including acoustic, prosodic and phonetic features. In order to integrate the cues to distinguish one language from another, we used a maximum a posteriori decision rule to connect an acoustic model, a duration model and a language model. In the experiments, we have achieved 34.5% (LID) and 17.7% (ASR) error rate reduction comparing with one stage LVCSR-based system.

Index Terms: Code-Switching Speech, Language Identification, Speech Recognition, Linguistic Cues

1. Introduction

The speech of code-switching is defined as using more than one language, variety, or style by a speaker within an utterance or discourse. To use code-switching speech is a common phenomenon in many bilingual/multilingual societies. For example, "我去 Starbucks 買一杯 coffee." ("I go to Starbucks to buy a cup of coffee.") is a typical Mandarin-English code-switching speech that we can hear in daily conversations. Such a speaking style could also be usually found in many bi-lingual or multilingual societies, such as French-German in Switzerland, English-Spanish in United States, Cantonese-English in Hong Kong [1] and Mandarin-Taiwanese in Taiwan [2].

A task of recognizing a code-switching utterance is full of challenges on the automatic speech recognition (ASR). Though multi-lingual speech recognizer can recognize several different spoken languages, the input test data has to contain only one language. Therefore, a speech recognition task on a code-switching speech, without the language boundaries and language identification information, does not perform well as that evaluating on a single language utterance [1-3]. According to this condition, we divided this task into two stages: one is language identification (LID) and the other is speech recognition. The language identifier offers the reliable language boundaries and language identification information. Then, based on the language boundary information, this task of recognizing code-switching utterances becomes that of recognizing bi-lingual utterances. Under this framework, the performance of the LID will directly influence the final speech recognition results. Therefore, to design a trustworthy LID system is an important research work.

During the past decades, a number of LID approaches for the multilingual speech were proposed, such as the parallel syllable-like unit selection [4], Gaussian mixture model tokenization [5] and large vocabulary continuous speech recognition (LVCSR) [6]. Of course, if the test utterance is longer, the performance will be better. However, the linguistic switching units in code-switching speech are smaller than a sentence. They are usually in clause or word levels. A Mandarin-Taiwanese code-switching example is shown in figure 1.

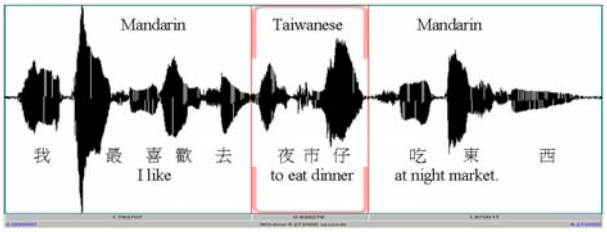

Figure 1. An example of Mandarin-Taiwanese code-switching utterance. The total number of the syllable is 11 and it contains 4.3 seconds. The language order is Mandarin-Taiwanese-Mandarin. The duration of the first Mandarin segment is 1.8 seconds, and Taiwanese duration only contains 0.8 second.

There are many cues that we could use to distinguish one language from another. According to [7], LID cues fall into five groups, acoustic, prosodic, phonetic, lexical and syntactic. In this paper, the proposed LID system integrates acoustic, prosodic and phonetic cues to evaluate on a Mandarin-Taiwanese code-switching speech. In order to obtain these cues, we firstly pass the utterance to a language-dependent LVCSR to get the phonetic cues, such as the information in the syllable (including lexical tone) level. Secondly, based on the recognized syllables, we used a quantized duration model to obtain the prosodic cues of each recognized syllables. Then we used a maximum a posteriori decision rule to integrate all the cues to train a syllable model, duration model and word bi-gram language model. Besides, the specific language categories and their corresponding boundaries in a code-switching utterance are determined by the Viterbi algorithm. Finally, based on the result of LID, a language dependent speech recognizer is used to detect the contents of the utterances.

2. Speech Corpus

The corpus we used is divided into three sets: two mono-lingual training data (Train-ML), and a code-switching test data (Test-CS). Beside, in order to evaluate how the proposed LID system performs on a single language speech, we segment the Test-CS manually and it is named Test-ML. The descriptions of the corpora are shown in the following:

- *Train-ML:* Two sets of mono-lingual (Taiwanese and Mandarin) training data. Each of the data sets included 100 speakers, and each of the speakers recorded 891 phonetically abundant utterances. This data set is used not only to train the acoustic models of LVCSR, but also to estimate the duration models of the tonal syllables.

- *Test-CS:* A Mandarin-Taiwanese code-switching testing set. In this set, another 24 speakers recorded 4600 code-switching utterances with total 4.8 hours. The set is designed that Mandarin is the matrix language and Taiwanese is embedded language. The grammatical structure is fundamentally based on Mandarin. The Taiwanese words are just used to substitute its Mandarin equivalent. Basically, each of the sentences has one or two Taiwanese words, and each of the words includes several tonal syllables. An example is shown in the figure 1.
- *Test-ML:* A mono-lingual test set extracted form Test-CS. We manually segmented and then extracted each of mono-lingual speech from Test-CS. Take figure 1 as an example, we extracted 3 mono-lingual speech segments into Test-ML. In that way, we can evaluate and compare Test-ML with Test-CS where the former is with the definite language boundaries but the latter is not. We totally extracted 6972 Mandarin and 3547 Taiwanese segments. The average duration of each extracted mono-lingual segment is around 1.5 seconds. All significant information of the corpus is shown in the Table 1.

	languages	number of speakers	number of utterances	ALPU (in sec.)	hours
Train-Bi	Mandarin	100	43,078	0.94	11.3
	Taiwanese	100	46,086	0.87	11.2
Test-CS	CS	24	4,600	3.79	4.8
Test-ML	Mandarin	24	6,150	1.70	2.9
	Taiwanese	24	4,987	1.37	1.9

Table 1. Statistics of the training and testing corpora, where Bi represents bi-lingual speech, CS represents code-switching utterances, ML means mono-language utterances and ALPU is the average length (in sec.) per utterance.

3. Baseline System

The technology of the parallel-based LVCSR with multiple languages assists LID systems to achieve high classification accuracy. In [6], the LID systems are built on language-dependent recognizers which include phoneme-based acoustic models and bi-gram or trigram phoneme language models. According to LVCSR process, the LID systems integrate the information of the higher linguistic knowledge and the recognizer results. Then, based on the integrated information, such as the highest likelihood, the LID systems decide a specific language given a test speech utterance of an unknown language. Nevertheless, most of the LVCSR-based approaches to LID systems are performed on a single language utterance. In this paper, we proposed a LVCSR-based LID system for the code-switching speech.

In this paper, we used the following procedures to deal with LID and ASR to take on code-switching utterances. Firstly, we extract the feature vectors, $\vec{O}_{cs}$, from the code-switching speech, $\vec{U}_{cs}$ in the proposed LVCSR-based LID system. Secondly, we build a language-dependent recognizer with Mandarin and Taiwanese acoustic models: AM$_M$ and AM$_T$. Orthographically, these two languages are monosyllabic languages. For this reason, we use the context-dependent Initial and Final as the model units. Initials and finals were adopted traditionally as the smallest natural pronunciation units in Mandarin and Taiwanese speech. They are a little different from phonemes in modern phonetics. The Initial corresponds to one phoneme, and the Final may consist of one or several phonemes. Then, in order to deal with code-switching speech, we use text corpora from both languages to train a union language model. The language mode is bi-gram. In this system, we added the language tag for each of the syllable in the text corpora. Therefore, according to the tag of the language of each recognized syllables, we can get the LID results.

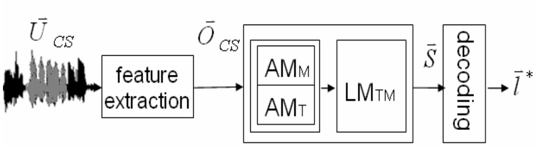

Figure 2. LVCSR-based LID system for code-switching utterances.

4. Cues of Mandarin and Taiwanese

There are a variety of cues that humans can use to distinguish one language from another. For example, different languages have different styles of acoustic articulation, prosody, phoneme sets, words, grammar and syntax. Based on this concept, recent studies have explored different types of speech cues which include acoustic parameters, prosody, phonetic and lexical knowledge to identify languages [6, 8].

In order to efficiently select the cues that are suitable to classify Mandarin and Taiwanese speech, we have to understand how they differ from each other. Basically, Taiwanese and Mandarin are the dialects of the Chinese languages, but they are mutually unintelligible [9]. Therefore, we should find the combination cues which can increase the probabilities of separating Mandarin and Taiwanese. Fortunately, based on linguistic literatures [10], we adopt and integrate the following cues.

The phonetic cue, such as the syllable, is the smallest meaningful linguistic unit. As mentioned before, the two languages are monosyllabic languages, and each of the language has its own syllable set. There are 408 syllables in Mandarin, and 709 syllables in Taiwanese. Discarding the overlapped part, 189 syllables, they remain 735 syllables which can be used to distinguish one from the other. The percentage of the overlapped part is 20%. In order to reduce the overlapped set of phonemes, we further use prosodic cues.

The prosodic cue, such as tone/fundamental frequency, is also a critical cue to discriminate Mandarin and Taiwanese. There are 4 and 7 lexical tones in Mandarin and Taiwanese, respectively. Different tones with the same syllable have distinguishable lexical meanings. There are 1288 and 2878 tonal syllables of Mandarin and Taiwanese, but only 647 out of 3519 tonal syllables are overlapped. The percentage of the overlapped part is deduced to 18%. Therefore, the percentage of the overlap phoneme decreases from base syllable to tonal syllable. That means if we add the tonal cues of the syllable, it essentially increases the probabilities of discriminating Mandarin from Taiwanese.

On the other hand, in order to distinguish the overlap set of tonal syllables, we use the rhythm cue which could be defined as the variation of the duration of a series of sounds. Rhythmic differences between languages are then mostly related to their syllable structure and the presence of vowel reduction [11]. In this paper, we used the cue to measure the average length of tonal syllables uttered per second that reveals the mean and variance of syllable duration. On the basis of the rhythm may vary due to the context in the different language with the same tonal syllable, we should imply the rhythm cue to distinguish Mandarin and Taiwanese.

5. Acoustic, Prosodic, and Phonetic LID System (APPLID-SYS)

5.1. System Overview

We integrate acoustic, phonetic and prosodic cues to generate a LID system for code-switching speech. The system is shown in figure 3: first one is a LVCSR and the second one is a integrated LID system including quantized acoustic, duration and language models, which we name as APPLID-SYS. The last one is a recognizer.

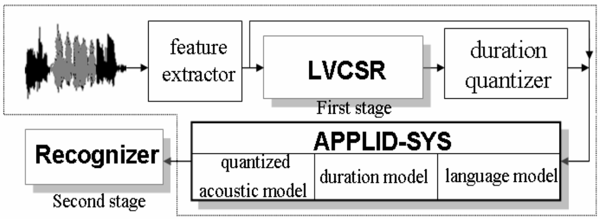

Figure 3. The procedure of two-stage code-switching speech recognition system with an APPLID-SYS and a recognizer.

The features include both acoustic and prosodic cues, o={o$_1$, …, o$_t$}, where *t* is a number of frame of a utterance. The outputs are 1) recognized tonal syllables, s={s$_1$, …, s$_n$}, where *n* the total numbers of tonal syllables, 2) corresponding duration of recognized tonal syllables, $d'_s \in R$, and 3) language tags of recognized tonal syllables, l$_s$={T, B, M}, where B represent the overlapped part of phonetic set, and T and M are the language dependent phonetic sets for Taiwanese and Mandarin.

In APPLID-SYS, the most likely language of each tonal syllable l$_s^*$={T, M} is decided by the following expression to formulize it.

$$l_s^* = \arg\max_{\forall l} P(l \mid o, s, d_s) \qquad (1)$$

where $d_s \in \{S, L\}$ is the quantized duration of each recognized tonal syllable. We apply the formula of [12] and rewrite as

$$l_s^* \cong \arg\max_{\forall l} p(o \mid d_s, l, s) p(d_s \mid l, s) p(s \mid l) p(l) \qquad (2)$$

The four probability expressions in (2) are organized as
1. *P(o|l,s,d$_s$)*, tonal syllable acoustic model.
2. *P(d$_s$|l,s)*, duration model.
3. *P(s|l)*, the tonal syllabic language model.
4. *P(l)*, the a priori probability for underlining language.

For practical reasons, the four components, *P(o|l,s,d$_s$)*, *P(d$_s$|l,s)*, *P(s|l)*, *and P(l)* are assumed to be independent of each other, and the weight of each probability is equal. We assume that *a* priori language probability for each language with code-switching speech is a uniform distribution. Therefore, the hypothesized language identity is then determined by maximizing the log-likelihood of language for each tonal syllable and is estimated as follows:

$$l_s^* \cong \arg\max_{\forall l} \{\log\ p(o \mid d_s, l, s) + \log\ p(d_s \mid l, s) + \log\ p(s \mid l) p(l)\} \qquad (3)$$

The quantized acoustic model measures the likelihood of different acoustic features in the syllabic elements with different languages. The duration model evaluates the probability of duration distributions of the syllabic elements across languages. The syllabic language model calculates the transformation or the probability of language switching by using the word-based bi-gram model.

5.2. Quantized Acoustic Modeling

The expression, *P(o|l,s,d$_s$)*, represents the conditional probability associated with the feature vector sequence *o* by giving language *l*, tonal syllable *s* and its quantized duration d_s. The original duration of each tonal syllable, d'_s, is a real number, but it is difficult to deal with if we train the duration as the parameters in the acoustic model. In order to simplify the number of parameters, we applied the concept of [13], and we quantized the duration *d* for each tonal syllable *s* into two levels: long and short. The procedures are demonstrated in the following steps:

● Step1: Forced alignment: In order to get the duration value of each tonal syllable, we used HMM-based forced-alignment approach to segment all tonal syllables in the training corpora for both languages.

● Step2: Average duration estimation: A histogram of duration for each tonal syllable emitted from forced alignment is collected and then we estimated the average duration. The average of the tonal syllables duration is 0.34 second form our analysis.

● Step3: Quantization: A -L suffix is appended to the tonal syllable if the duration is longer than the average duration, otherwise, we use a -S suffix.

An example is shown in figure 4. The input duration of the syllables /kin1T/ and /ten1T/ are 0.24 and 0.57 second. The average duration (threshold) of the syllables /kin1T/, and /ten1T/ are 0.27, and 0.31 second, respectively. Thus, these two input syllables are appended with –S and –L and they become /kin1T-S/ and /ten1T-L/.

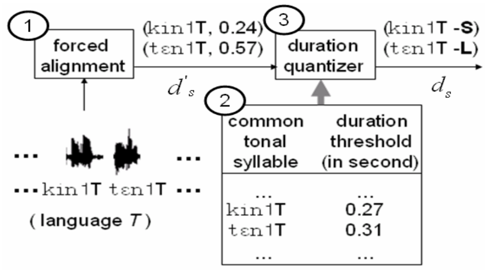

Figure 4 An example of language dependent quantized duration for each tonal syllable.

5.3. Duration Modeling and Language Modeling

The expression, *P(d$_s$|l,s),* represents the conditional probability associated with quantized duration *d* given the tonal syllable *s* and language *l*. The model reflects the rhythm cue in the syllabic level of different languages. Because the parameters related to rhythm is based on the information of syllable timing or syllable duration [14]. Therefore, for each quantized tonal syllable in each language, we use a statistic approach to train the probabilistic-based duration model from a bilingual corpus.

The expression *P(s|l)* represents the conditional probability associated with tonal syllable *s* given the language *l*, and it refers to the language model. In this part, we employed a word-based bi-gram statistical framework.

6. Experiment and Conclusion

6.1. LVCSR-based Results

In experiment parameter settings, firstly, for the feature extraction, we used 13-MFCCs and fundamental frequency with their first and second differential values, thus the total dimensions of the features are 42. Secondly, a HMM-based recognizer is used with tri-phone acoustic models and with a word-based bi-gram language model trained from 20,000 words code-switching text data. The complexity of the language model is 245. On the other hand, due to there are only two languages within *Test-CS* set, the measurement of the LID results is adapted [14] method. This method not only considers the target category but also take the time frame of the target into account. The error rates are defined as follows:

LID error rate = (# of FAs + # of FRs) / (total # of labels)
where FA rate = (# of FAs) / (total # of non-target), and
FR rate = (# of FRs) / (total # of targets)

For measuring the *Test-ML*, we used voting algorithm, which decides the segment belonging which language dependents on the most syllables belong which language within a segment. Besides, the speech recognition measurement is: ASR error rate = 1- syllable accuracy rate.

The results of baseline system are shown in table 2. In general, because of known language boundaries, the LID performance of the *Test-ML* is better than that of *Test-CS*. The similar tendency is in ASR performance.

Baseline System	LID error rate	ASR error rate
Test-ML (Mandarin)	15.5%	39.2%
Test-ML (Taiwanese)	17.8%	40.6%
Test-CS	28.1%	44.7%

Table 2. The LID and ASR error rates of baseline system.

6.2. APPLID-SYS Results

Based on equation 2, we evaluated each cue independently, and set three conditions.

- *A*: we only used quantized acoustic model. It means that the duration and language models are set to be uniform distributions;
- *A+P*: both quantized acoustic model and duration model were used. It means that only the language model is set to be uniform distribution.
- *A+P+P*: all models were used.

The LID and ASR results are shown in table 3 and 4. Comparing the second column of table 3 with that of table 2, we gained 15.7% error rate reduction of Test-CS and average gained 14.7% error rate reduction of *Test-ML*. Besides, as we added the duration and language model, the LID performance of using APPLID-SYS improved more than LVESR-based system. We got the best error reduction rate, 34.5% (from 28.1% to 18.4%), if we used APPLID-SYS comparing with baseline system.

The second stage recognizer is a language dependent system, and the setting of the feature extraction and the acoustic model is the same of baseline system. However, the language model is bi-gram language dependent framework. The complexity of language model for Mandarin and Taiwanese is 64 and 87. The text corpus for both languages contains 10,000 vocabularies.

For ASR results, the error reduction is not as much as the LID performance when we sequentially add duration model and language model. For the Test-CS set, the error reduction rate is 5.9% if we used A to A+P+P. However, based on our two stage approach, the final ASR performance is better than one stage baseline system. We have achieved 17.7% (from 44.7% to 36.8%) error rate reduction on *Test-CS*.

APPLID-SYS	A	A+P	A+P+P
Test-ML (Mandarin)	14.2%	11.5%	9.9%
Test-ML (Taiwanese)	13.8%	11.7%	11.2%
Test-CS	23.7%	19.6%	18.4%

Table 3 The LID error rates of APPLID-SYS.

ASR	A	A+P	A+P+P
Test-ML (Mandarin)	35.3%	34.1%	33.5%
Test-ML (Taiwanese)	38.2%	35.9%	34.3%
Test-CS	39.1%	37.2%	36.8%

Table 4 The syllable error rate of ASR

In this paper, we proposed a two-stage system, language identifier and speech recognizer, and evaluated on Mandarin-Taiwanese code-switching speech. The language identifier integrated multiple level linguistic cues to distinguish one language form another. The results also show that the proposed system significantly contributed to error reduction for both LID and ASR comparing with one stage LVCSR-based system.

7. References

[1] Y.C. Chan, P.C. Ching, Tan Lee and Houwei Cao "Automatic speech recognition of Cantonese-English Code-Mixing utterances," In Proc. of ICSLP, 2006.

[2] Dau-Cheng Lyu, Ren-Yuan Lyu, Yuang-chin Chiang and Chun-Nan Hsu, "Speech Recognition on Code-switching Among the Chinese Dialects," In Proc. of ICASSP, 2006.

[3] Chung-Hsien Wu, Yu-Hsien Chiu, Chi-Jiun Shia, and Chun-Yu Lin "Automatic Segmentation and Identification of Mixed-Language Speech Using Delta-BIC and LSA-Based GMMs," IEEE Transactions On Audio, Speech, And Language Processing, 14-1, Jan. 2006.

[4] Pedro A. Torres-Carrasquillo, Douglas A. Reynolds. and Deller Jr., "Language identification using Gaussian. Mixture Model Tokenization," In Proc. of ICASSP, 2002

[5] T. Nagarajan and Hema A. Murthy, "Language Identification Using Parallel Syllable-Like Unit Recognition," In Proc. of ICASSP, 2004.

[6] Schultz, T., Rogina, I., Waibel, A., "LVCSR-based Language Identification," In Proc. of ICASSP, 1996

[7] R. Tong, B. Ma, D. Zhu, H. Li, E. S. Chng, "Integrating Acoustic, Prosodic and Phonotactic Features for Spoken Language Identification," In Proc. of ICASSP 2006

[8] Rouas, J.-L. Farinas, J. Pellegrino, and F. Andre-Obrecht, "Modeling prosody for. language identification on read and spontaneous speech," In Proc. of ICASSP, 2003

[9] R.Y. Lyu, et al., "A Unified Framework for Large Vocabulary Speech Recognition of Mutually Unintelligible Chinese "Regionalects"," In Proc. of ICSLP, 2004

[10] Cheng, Robert L. "Taiwanese and Mandarin Structures and Their Developmental Trends in Taiwan, " Book I-IV, Taipei: Yuan-Liou Publishing Co., Ltd., 1997

[11] Farinas, J., Pellegrino, F., Rouas, J.L., Andre'-Obrecht, R., "Merging segmental and rhythmic features for automatic language identification," In Proc. of ICASSP, 2002

[12] Hazen, T. J. and Zue, V. W. “Segment-based. automatic language identification”, Journal of. the Acoustical Society of America, 101(4), 2323-2331, 1997

[13] Zissman, M.A. "Language Identification Using Phoneme Recognition and Phonotacticlanguage Modeling," In Proc. of ICASSP, 1995

[14] J. Li and C.-H. Lee, "On Designing and Evaluating Speech. Event Detectors," In Proc. of InterSpeech, Lisbon, Sept. 2005