Language Identification by Using Syllable-based Duration Classification on Code-switching Speech

Dau-cheng Lyu[2,3], Ren-yuan Lyu[1], Yuang-chin Chiang[4], Chun-nan Hsu[3],

[1] Dept. of Computer Science and Information Engineering, Chang Gung University, Taiwan
[2] Dept. of Electrical Engineering, Chang Gung University, Taiwan
[3] Institute of Information Science, Academia Sinica, Taiwan
[4] Institute of statistics, National Tsing Hua University, Taiwan
renyuan.lyu@gmail.com

Abstract. Many approaches to automatic spoken language identification (LID) on monolingual speech are successfully, but LID on the code-switching speech identifying at least 2 languages from one acoustic utterance challenges these approaches. In [6], we have successfully used one-pass approach to recognize the Chinese character on the Mandarin-Taiwanese code-switching speech. In this paper, we introduce a classification method (named syllable-based duration classification) based on three clues: recognized common tonal syllable tonal syllable, the corresponding duration and speech signal to identify specific language from code-switching speech. Experimental results show that the performance of the proposed LID approach on code-switching speech exhibits closely to that of parallel tonal syllable recognition LID system on monolingual speech.

Keywords: language identification, code-switching speech.

1 Introduction

Code-switching is defined as the use of more than one language, variety, or style by a speaker within an utterance or discourse. It is a common phenomenon in many bilingual societies. In Taiwan, at least two languages (or dialects, as some linguists prefer to call them) - Mandarin and Taiwanese- are frequently mixed and spoken in daily conversations.

For the monolingual LID system development, the parallel syllable recognition (PSR) was adopted, which is similar to the method of parallel phone recognition (PPR), and this approach is widely used in the automatic LID researches. [1,-5] Here, the reason to use syllable as the recognized result instead of phone is because both Taiwanese and Mandarin are syllabic languages. Another approach, which is called parallel phone recognition followed by language modeling (parallel PRLM), used language-dependent acoustic phone models to convert speech utterances into sequences of phone symbols with language decoding followed. After that, these acoustic and language scores are combined into language-specific scores for making an LID decision. Compared with parallel PRLM, PSR uses integrated acoustic models

to allow the syllable recognizer to use the language-specific syllabic constraints during decoding process, and it is better than applying those constraints after syllable recognition. The most likely syllable sequence identified during recognition is optimal with respect to some combination of both the acoustics and linguistics.

However, all these approaches were confronted with an apparent difficulty. That is, they use speech signal length in sentence level or 10-45 seconds as test speech and then the language is decided by which gets maximum number of unique phonetic unit is noted as the winner for the test utterance. In our case, code-switching speech, the length of language changing may be intra-sentence or word-based level, and we can not identify the language using above approach, because there may have at least two languages embedded in a test utterance. Therefore, we have to decide the language identity in a very short time of speech utterance.

In this paper, we propose an alternative to deal with the code-switching speech LID task, which is SBDC (syllable-based duration classification). This framework could identify the language in syllable level which avoids the shortcoming of LID system in sentence or utterance-based utterances. Besides, to identify a language in each syllable that performs more precise language boundary in the code-switching speech. In this framework, we, firstly, extract acoustic and pitch features from code-switching utterance, secondly, the features are recognized as tonal syllable by our pervious recognizer [6]. Thirdly, by given the tonal syllable and its duration information, we use SBDC to identify the language for each common tonal syllable. Finally, the language smoother modifies the language identify in a statistical approach form training a code-switching speech corpus.

The structure of the paper is as follows: A LVCSR-based LID system is introduced in Section 2. The phonetic characteristic between Mandarin and Taiwanese is introduced in Section 3. In Section 4 the SBDC-based LID system is described. Finally, the performed experiments and achieved results are presented.

2 LVCSR-based LID

It is known that LVCSR-based systems achieve high performance in language identification since they use knowledge from phoneme and phoneme sequence to word and word sequence. In [7], the LVCSR-based systems were shown to perform well in language identification. Unlike mono-lingual speech LID system [8], we implement a multi-lingual LVCSR-based system [9] as our code-switching speech LID baseline system. Fig 1 shows a block diagram of the system which includes two recognizers and each recognizer contains its won acoustic model and language model, such as AM_T, LM_T.

In this paper, the multi-lingual LVCSR-based system requires significant tonal syllable level transcribed Mandarin and Taiwanese speech data for training the acoustic and language models. During the test phase, the recognizer is employed a unified approach to recognize each tonal syllable. The step of decoding translates each tonal syllable to its won language by phonetic knowledge. It is among the most computationally complex algorithms and achieves very high language identification accuracy.

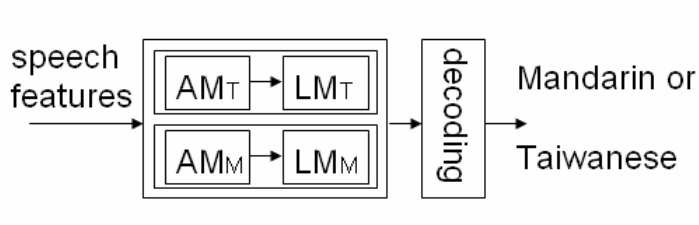

Fig. 1. A diagram of the unified LVCSR-based LID architecture for code-switching speech.

3 Linguistic Characteristics of Mandarin and Taiwanese

In order to achieve reasonable identification accuracy in Taiwanese and Mandarin identification, it is necessary to understand how languages differ. They differ in the following ways:

1. Phonemic System:
2. Tone (e.g., Mandarin has four tones , Taiwanese has seven tones)
3. Tonal Syllable (there are 677 tonal syllables only belonging to Mandarin, 2231 ones only belonging to Taiwanese, and 610 tonal syllables exist in both languages by using IPA notation)
4. Lexical distribution
5. Rhythmical characteristics
6. Pace (average number of tonal syllables uttered per second) or tonal syllable duration
7. Intonation or lexical stress

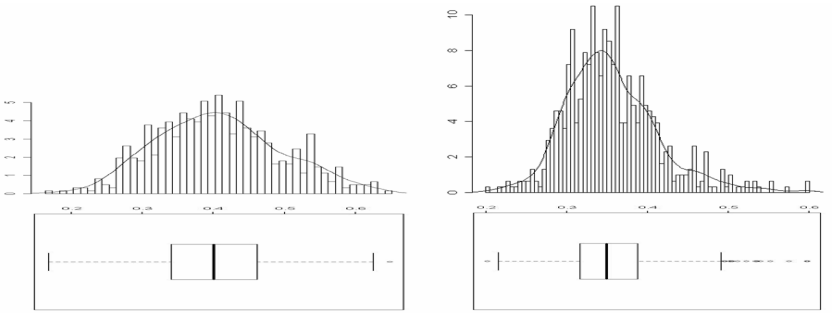

Fig. 2. The average duration distribution of common tonal syllables in Taiwanese (left) and Mandarin (right). The x-axle is the duration in second unit and y-axle is the number of common tonal syllable. (610 is the total value if the summing all y-axle value in x-axle)

The duration distribution of common tonal syllables (610) estimating from training corpus is shown in Fig 2, and they have different mean and variation. The numbers of the total sample for Mandarin are 54371 samples and 39589 samples for Taiwanese. The syllable duration of Taiwanese is about 04 sec. and 0.3 sec. for Mandarin syllable.

4 SBDC LID

From the analysis of sec 3, we have an idea to discriminate Taiwanese from Mandarin by the duration discrepancy of the common tonal syllables. Thus, in this section, we develop a new approach to identify language for each tonal common syllable on code-switching speech.

4.1 System Overview

There are five components, including a feature extractor, a unified speech recognizer, a common tonal syllable extractor and a language smoother, in our code-switching LID system. Figure 3 illustrates the process, and the procedures are as the followings:

1) The code-switching speech input utterance is pre-extracted into a sequence of MFCCs-based feature vectors $O_T = (o_1, o_2, ..., o_T)$ with the length (frame number) *T*.
2) The unified speech recognizer [6] receives the features as the input and finds the best hypothesis tonal syllable $S^N = (s^1, s^2, ..., s^N)$ and its corresponding duration D_R, where *N* is the distinct tonal syllables for all the languages and *R* is the real number which represents the duration of each hypothesis tonal syllable. According to the pronunciation dictionary, each of the hypothesis tonal syllable is further represented by the language code $L = \{m, t, c\}$, where *m* represents Mandarin, *t* represents Taiwanese and *c* means common language. The tonal syllables with the common language mean that they exist in both Mandarin and Taiwanese, and this kind of phenomenon is caused by the union phonetic representation of the unified speech recognizer. An example is shown in the figure 4.
3) We only extract the speech segment with common language, S_c, for discriminating between Mandarin and Taiwanese.
4) The three parameters, O_T, S_c and D_R are as the inputs to train the syllable-based duration classifier (SBDC). The output is language specific tonal syllable, S_{ct} and S_{cm} for instance. This part will describe particularly in the section 4.2.
5) In practice, the unit of code-switching language appears as a word whose unit exceeds duration. Under this assumption, the smoothing process is involved to eliminate the unreasonable language switching with a short interval by the language modeling after joining the parts of S_{cm}, S_{ct} and S_t, S_m. The final output is $S_{\tilde{m}}$ or $S_{\tilde{t}}$ which is a tonal syllable with the language identity of Mandarin or Taiwanese.

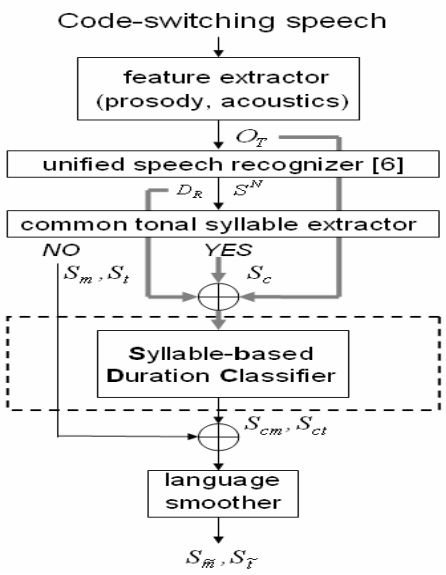

Fig. 3. The flow chart of SBDC LID system

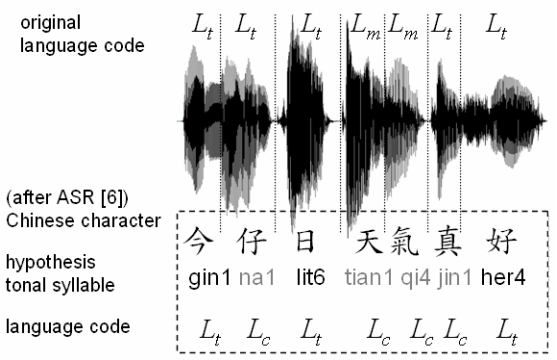

Fig. 4. Example of language code.

4.2 Probabilistic Framework

The most likely language L_i by given three parameters: the acoustic information O_T, the common tonal syllable S_c, and its duration D_R, is found using the following expression:

$$Li(O_T) = \arg\max_i P(L_i \mid O_T, S_c, D_R) \quad (1)$$

Using standard probability theory, this expression can be equivalently written as

$$L_i(O_T) = \arg\max_i P(O_T \mid L_i, S_c, D_R) P(D_R \mid L_i, S_c) P(S_c \mid L_i) P(L_i) \qquad (2)$$

The four probability expressions in (2) are organized in such a way that duration and common tonal syllable information are contained in separate terms. In modeling, these terms become known as

1. $P(O_T \mid L_i, S_c, D_R)$ Common tonal syllable acoustic model.
2. $P(D_R \mid L_i, S_c)$ Duration model.
3. $P(S_c \mid L_i)$ The phonetic language model.
4. $P(L_i)$ The a priori language probability.

Assuming that a priori language probability for each language on code-switching speech is equal, and phonetic language model for common hypothesis tonal syllable is also equal. The hypothesized language is determined by maximizing the log-likelihood of language L_i with respect to speech O_T and is estimated as follows:

$$L_i(O_T) = \arg\max_i \{\log P(O_T \mid L_i, S_c, D_R) + \log P(D_R \mid L_i, S_c)\} \qquad (3)$$

According to [3], the syllabic information is contained in two separate models: the syllabic acoustic model and the syllabic duration model, which are shown in Fig 5. In subsequent sections these models will simply be referred to as the acoustic model and the duration model. The acoustic model accounts for the different acoustic realizations of the syllabic elements that may occur across languages, whereas the duration model accounts for the probability distributions of the syllabic elements, and captures the differences that can occur in duration structures of different languages due to the boundary or segmented created by variations in the common tonal syllabic durations. This organization provides a useful structure for evaluating the relative contribution towards language identification that acoustic and duration information provide.

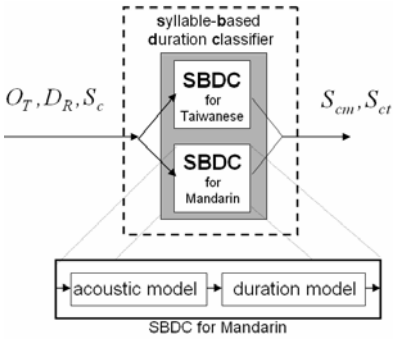

Fig. 5. Illustration of syllable-based duration classifier component.

4.3 Acoustic Model of SBDC

The expression $P(O_T \mid S_c, D_R, L_i)$ is called the acoustic model, which is used to capture information about the acoustic realizations of each of the common tonal

syllable used in each language. However, the duration of each common tonal syllable is a real number form 0 to 1 and that is parametric difficultly. To simplify the parameter of the acoustic model, like the idea of [1], the duration *R* of the D_R for each common tonal syllable S_c is quantized into two levels: long and short, by the following steps:

Step1: Forced alignment:

In the training phase, we need to get the duration for each tonal syllable, because our transcription of the training speech only contains the pronunciation, not including duration information. Therefore, we used HMM-based method to get the duration value of each tonal syllable by forced alignment approach on training corpora for both languages.

Step2: Average duration estimation:

A histogram of duration for each tonal syllable emitted from forced alignment is collected and the average duration determined.

Step3: Quantization:

A –L suffix is appended to all tonal syllables having duration longer than the average duration for that tonal syllable, and –S suffix is appended to all tonal syllables having duration shorter than the average duration for that tonal syllable. The diagram for these steps for language *i* is illustrated in the Fig. 6

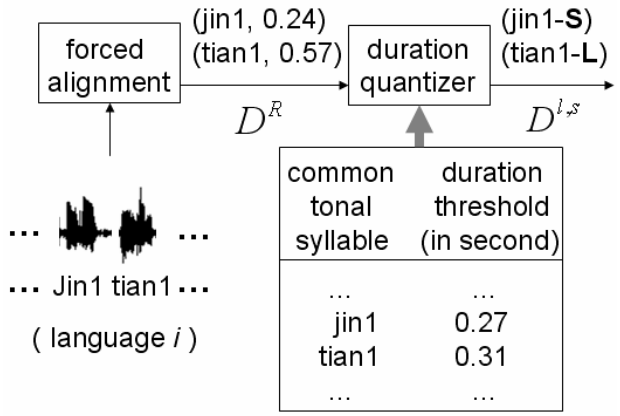

Fig. 6. An example of tagging quantized duration for each common tonal syllable. (Form D_R to D_l or D_s)

4.4 Duration Model of SBDC

The expression $P(D_R \mid L_i, S_c)$ captures the segment duration information in a common tonal syllable. While there may be very useful information to separate the difference between Mandarin and Taiwanese in syllable level. The probability can be modeled with a mixture of Gaussian models. The Gaussians in each mixture are then iteratively re-estimated to maximize the average likelihood score of the vectors in the

training data. To ensure proper amounts of training data for each mixture of Gaussians, the number of Gaussian used to model each syllable in each language is determined by the amount of the training data.

4.5 Language Smoother (LS)

The goal of the language smoother is to modify the language identity to be more reasonable in language switching by an N-gram language model trained from a real code-switching corpus. An example is shown in Fig 7.

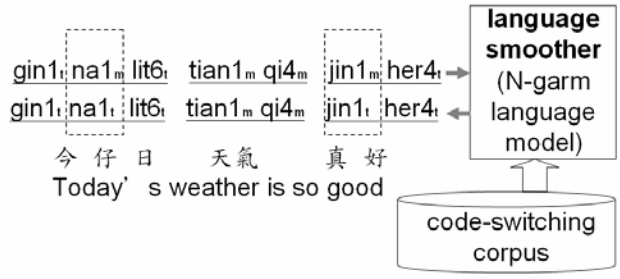

Fig. 7. An example for merging language identification results by language smoother.

5 Experiments and Results

The goal of the experiment is to verify that SBDC-based system could have high accuracy syllable LID rate and to outperform a LVCSR-based system. In addition, we also evaluate the performance of our proposed approach is close to that on monolingual speech which maybe the upper bound performance on code-switching speech.

5.1 Corpus and Experiment setup

The speech corpus used in all the experiments were divided into three parts, namely, the training set, evaluating set and the testing set. The training set consists of two mono-lingual Taiwanese and Mandarin speech data, which includes 100 speakers. Each speaker read about 700 phonetically abundant utterances in both languages. The evaluating set is to train the back-off bi-gram code-switching language model, which estimates the probability of language translation, and adapts the threshold of syllable duration quantizer. For testing data set, another 12 speakers were asked to record 3000 Mandarin-Taiwanese code-switching utterances. Among these utterances, at least one Taiwanese word is embedded into a Mandarin carrier sentence. The length of each word is various from one to eight syllables. The statistics of the corpus used here are listed in Table 1.

The acoustic features used in SBDC are the same with in [5], they are: mel-frequency cepstral coefficients (MFCC) which includes 12 cepstral coefficients,

normalized energy and prosody information. The first and second derivatives of parameters are also included. The acoustic model of SBDC is used HMM-based approach in tonal syllable unit with duration tag for both languages, and each HMM has seven states. We have 2440 (610 common tonal syllables, and each one has two duration classes and two languages) HMM, and the final number of the mixture in each state dependents on the occurrence of training data. The more the training data of the state has, the more of the mixture number is.

Table 1. Statistics of the bi-lingual speech corpus used for training and testing sets. M: Mandarin, T: Taiwanese, CS: code-switching utterances

	Language	No. of Speakers	No. of Syllable	No. of Hours
Training set	M	100	112,032	11.3
	T	100	124,768	11.2
Evaluating set	CS	4	12,680	1.10
Test set	CS.	12	41,681	3.31

5.2 Results

We compare the LID performance on two different types of speech: f monolingual speech, and on code-switching speech. The contents of these two sets are the same, because we used manual segment to extract the part of monolingual speech form code-switching speech. For the monolingual speech, we used parallel tonal syllable recognition approach [5], and the approach is similar with PPR [1]. On the other hands, for the code-switching speech, we used three approaches which are LVCSR-based approach, SBDC and SBDC+LS. The last one is SBDC approach adding the language smoother. The 10K and 20K vocabulary size of SBDC and SBDC+LS approaches are followed the experimental condition of [6], because, before do SBDC approach, we need the recognized tonal syllable form unified ASR in [6]. The results are listed in the table2.

The experimental result for the monolingual speech has the highest LID accuracy rate, 88.05%, and this is the experimental upper bound performance of LID on code-switching speech using our method.

On the code-switching speech, using SBDC+LS approach has the best performance which is close to that in the monolingual speech. Both of the performances of using SBDC approach are more outstanding than LVCSR-based approach, the critical reason is according to the useful information such as duration of tonal syllable used during decision process.

Table 2. The LID accuracy rate for different approaches

	LID accuracy rate (%)	
monolingual speech	**88.05**	
code-switching speech	10K	20K
LVCSR-based	82.14	81.93
SBDC	86.08	84.78
SBDC+LS	**87.53**	**85.91**

6 Conclusion

In this paper, we used three clues: recognized common tonal syllable tonal syllable, the corresponding duration and speech signal, building a SBDC LID system to identify specific language from code-switching speech. The system's architecture is factorized as HMM-based quantized duration acoustic model in tonal syllable, GMM-based duration model and language smoother. The experimental results show a promising performance on LID accuracy rate to compare with the LVCSR-based system and the performance also approaches that in monolingual speech by using PSR method.

References

1. Zissman, M. A. "Comparison of four Applications to Automatic Language Identification of Telephone Speech," IEEE Trans. on Speech and Audio Proc., Vol. 4, No. 1, pp. 31-44, 1996
2. T. Nagarajan and Hema A. Murthy, "Language Identification Using Parallel Syllable-Like Unit Recognition," ICASSP, 2004
3. Hazen, T. J., & Zue, V. W., "Segment-Based Automatic Language Identification," Journal of Acoustic Society of America, April 1997.
4. Rongqing Huang, John H.L. Hansen, "DIALECT/ACCENT CLASSIFICATION VIA BOOSTED WORD MODELING," ICASSP 2005
5. Pedro A. Torres-Carrasquillo, Douglas A. Reynolds, J. R. Deller, Jr., "Language Identification Using Gaussian Mixture Model Tokenization," Proc. ICASSP 2002, pp. I-757-760.
6. Dau-Cheng Lyu, Ren-Yuan Lyu, Yuang-chin Chiang and Chun-Nan Hsu, "Speech Recognition on Code-Switching Among the Chinese Dialects," ICASSP, 2006
7. T. Schultz et al., "LVCSR-based Language Identification," Proc. ICASSP, pp. 781-784, Altlanta 1996.
8. J.L.Hieronymus, S.Kadambe, "Robust Spoken Language Identification using Large Vocabulary Speech Recognition", ICASSP, vol.2, pp.1111-1114, Munich, Germany, Apr., 1997
9. Santhosh C. Kumar, VP Mohandas and Haizhou Li, "Multilingual Speech Recognition: A Unified Approach", InterSpeech 2005 - Eurospeech - 9th European Conference on Speech Communication and Technology, September 4-8, 2005, Lisboa, Portugal