SPEECH RECOGNITION ON CODE-SWITCHING AMONG THE CHINESE DIALECTS

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
3Institute of Information Science, Academia Sinica, Taiwan
4 Institute of statistics, National Tsing Hua University, Taiwan
rylyu@mail.cgu.edu.tw; renyuan.lyu@gmail.com; http://msp.csie.cgu.edu.tw

ABSTRACT

We propose an integrated approach to do automatic speech recognition on code-switching utterances, where speakers switch back and forth between at least 2 languages. This one-pass framework avoids the degradation of accuracy due to the imperfectly intermediate decisions of language detection and language identification. It is based on a three-layer recognition scheme, which consists of a mixed-language HMM-based acoustic model, a knowledge-based plus data-driven probabilistic pronunciation model, and a tree-structured searching net. The traditional multi-pass recognizer including language boundary detection, language identification and language-dependent speech recognition is also implemented for comparison. Experimental results show that the proposed approach, with a much simpler recognition scheme, could achieve as high accuracy as that could be achieved by using the traditional approach.

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. [1][2] 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. [3] It also becomes a type of skilled performance in a public speech. Take a famous speaker's speech on a TV program for example, which was recorded, plot, and transcribed in 3 layers of labels as shown in figure 1, where we could find many instances of code-switching. The 3 layers of transcription labels include the Chinese character sequence, its corresponding meaning translated in English, and the language identification for Mandarin or Taiwanese. In this case, the speaker makes code-switching for about 5 times during a 15-second utterance, where there are totally 63 syllables. In this figure, we can find a significant break between block 1 and block 2 where it is obviously defined as the sentence boundary. Each block can be looked upon as a complete sentence. A language change between block 1 and block 2 is called an inter-sentential code-switching. On the other hand, the language change inside a sentence such as that occurs in block 2 is defined as an intra-sentential code-switching. This example just reflects the fact that code-switching has become more or less to the norm of the language community in Taiwan.

Our goal in this research is to develop an ASR system that could recognize a Mandarin-Taiwanese code-switching utterance as a Chinese character sequence. Recognizing an utterance in mixed languages has still been a challenge for the present ASR systems. An apparent approach to do speech recognition on this task could be described in the following; firstly, the system should be able to do language boundary detection [4] which segments the code-switching utterances into several mono-language speech segments. Secondly, language identification (LID) [5] system identities each speech segments as a specific language. Then we can recognize the specific mono-lingual utterance by language-dependent speech recognizers. We will implement this approach on code-switching utterances as baseline and call it as the multi-pass recognizer. That is, a multi-pass recognizer contains a language boundary detection module, a language identification module, and a mono-lingual ASR, which will be described shortly in the following section.

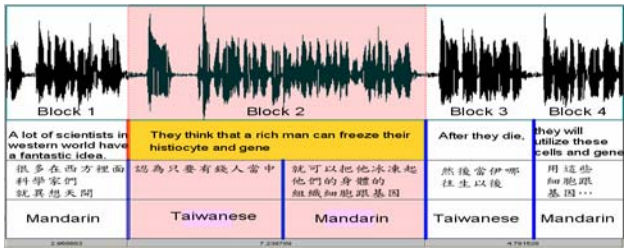

Figure 1. Example for Mandarin-Taiwanese code-switching speech is extracted from a TV program. This speech waveform is labeled in three layers, which represent the English translation, the Chinese character sequence and language identity, respectively.

Automatic language identification and language boundary detection have been studied by many researchers. A survey report had compared many the state-of-the-art of LID approaches and performance on the utterance with a single language. [6] Other reports did the language boundary detection and identification tasks on the mix-languages. One of them is detecting 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. [4] Another paper reported that the use of LSA-based 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]

However, all these approaches were confronted with an apparent difficulty. That is, the upper bound in performance of one stage of processing will be restricted to that of the previous stage. Therefore, the performance of language boundary detection will directly influence the results of language identification, and finally, that will limit the speech recognition performance.

In this paper, we propose an alternative to deal with the code-switching speech recognition task, which is called a one-pass framework for ASR system. This framework could avoid the performance loss in each processing stage of the above approach. In this framework, an utterance in mixed-languages is unnecessary to be segmented into short segments with only a mono-lingual speech. It is also unnecessary to do the language identification in the middle way of the speech recognition process. An integrated approach was thus proposed, which includes a union set of acoustic models, a bi-lingual pronunciation model, and a Chinese character-based continuous tree-structured searching net.

The structure of the paper is as follows: A traditional multi-pass scheme includes the LID systems on single and code-switching utterances are introduced in Section 2. The new proposed one-pass speech recognizer is described in Section 3. In Section 4 the performed experiments and achieved results are presented. Finally, we draw some conclusions in Section 5.

2. MUTI-PASS SPEECH RECOGNITION FOR CODESWITCHING UTTERANCES

For the speech recognition on code-switching utterances, the traditional multi-pass scheme could be shown in figure2. Before doing the language-dependent speech recognition, the code-switching utterances have to be processed via the language boundary detection and language identification (LID). Therefore, in the following, two approaches of automatic language identification frameworks on single language and code-switching utterances are introduced.

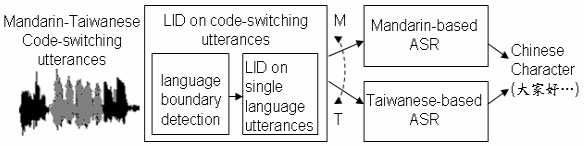

Figure 2. The diagram of ASR for code-switching speech in a multi-pass approach.

2.1. LID on single language utterances

For the 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. [6] 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.

To further improve the performance, other information, such as articulatory, acoustic and prosodic features have also been integrated into an LID system. Mandarin and Taiwanese are also tonal languages, using rhythmic and intonation features can be looked upon as efficient cues for discriminating languages.

In the PSR approach, the utterances are identified to language names by the higher likelihood scores emanating from language-dependent syllable recognizer. By allowing the syllable recognizer to use the Viterbi decoding rather 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 syllables..

2.2. LID on mixed-language utterances

Identifying mixed-languages in an utterance challenges the present LID system as described in the above (section 2.1). Segmenting such an utterance into two segments of different languages is crucial to the development of a LID system. For the mixed-language LID system proposed in this paper, we used the confidence scores to decide which region in one utterance belongs to its language by the recognized tonal syllables. This approach is similar to that mentioned in [4], which is used to detect the boundaries from English and Cantonese code-switching utterances by using bi-phone probabilities, and measuring the confidence that the recognized phones are in Cantonese. In our proposed approach, the confidence score is measured by likelihood of tonal syllables with the two languages during the same position of the recognized syllable sequences. For example, if a tonal syllable belongs to Taiwanese among a sequence of the results after decoding, then the boundary of the tonal syllable is fixed and the language inside the region of the tonal syllable is also identified. In order to train the language-specific acoustic models, each of the tonal syllables are tagged with the language information, and the acoustic models were trained according to their language specification.

Besides, two main innovations are further used in this task for capturing the characteristic of the mixed-languages. Firstly, we add the prosody information into the feature vectors because many literatures have demonstrate the fact that prosodic features are very helpful to distinguish tonal languages [11]. The prosody can be considered in two phases, representing the phrase accentuation and the local accentuation, as in Fujisaki's work. [12]. Secondly, we used bi-syllable likelihood of both languages as confidence measurement for syllable-based language identification. These two innovations are due to the fact that both the languages (Mandarin and Taiwanese) considered here are tonal and syllabic languages.

3. ONE-PASS RECOGNIZER

It is known that, all of the spoken varieties of the Chinese languages share a common formal written language, if we ignore the difference between traditional and simplified orthographies adopted in Taiwan and mainland China, respectively. In Taiwan, there were 85% commonly shared lexicon items between Mandarin and Taiwanese, although a number of special characters which are unique to Taiwanese are sometimes used in informal writing. [3] Because of these special linguistic characteristics, we have proposed a Chinese character-based one-pass recognition scheme, which has been proven successful in dealing with multiple dialects of the Chinese language using a unified framework. [13]

Unlike some conventional approaches, which divide the recognition task into language boundary detection and language identification, our proposed approach adopts a one-stage searching strategy, as shown in Figure 3. This approach is not only simpler than traditional multi-pass one, but also avoids the loss of hard decision in each stage, such as language boundary detection or LID, and it becomes easy to train and use integrated acoustic/pronunciation models.

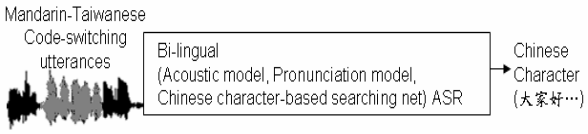

Figure 3. The diagram of ASR for code-switching speech in a one-pass approach

In this framework, Chinese character-based decoding can be implemented by searching in a three-layer network composed of an acoustic model layer, a lexical layer, and a grammar layer. There are at least 2 critical differences between our framework and the conventional one. 1) In the lexicon layer, character-to-pronunciation mapping can easily incorporate multiple pronunciations in multiple languages, including Japanese, Korean, and even Vietnamese, which also use Chinese characters. 2) In the grammar layer,

characters instead of syllables are used as nodes in the searching net. Under this ASR structure, we do not care which language the user speaks. No matter whether the language is Taiwanese, Mandarin or a mixture of them in one sentence, the ASR outputs the Chinese characters only. This makes it language/dialect independent.

In the acoustic modeling, we use international phonetic alphabet (IPA) to transcribe the corpus of the two languages discussed here. [14] Table 1 shows the statistical information of the phonemic inventory in different phonetic levels for Mandarin and Taiwanese. Sounds in different languages transcribed in the same phonemic symbols of IPA share the same speech material. Combining two languages in this manner reduces the number of syllables by 21%. In order to easily integrate tone information, we used the context-dependent Initial and tonal Final as acoustic units, and trained these models by sharing the data which belong to the same acoustic unit. Then, a divisive clustering algorithm was used to create context querying decision trees using four question sets, including an Initial set, a tonal Final set, the set of language properties, and a tonal information set.

	M	T	M∪T	M∩T
N_S	408	709	925	192(21%)
Tone	5	7	9	3
N_{TS}	1288	2878	3519	647(18%)
N_I	17	19	22	14(63%)
N_{TF}	295	225	416	104(25%)
N_{CDIF}	1656	3496	4374	778(18%)

Table 1. The statistic information of all Mandarin (M) and Taiwanese (T) linguistic units in four levels: the number of Syllable(N_S), the numbers of Tonal Syllables (N_{TS}), Initials (N_I), Tonal Finals (N_{TF}), and context-dependent Initial/tonal Finals (N_{CDIF}). ∩and∪ mean intersection and union, respectively.

Many Chinese characters are homographs which have multiple meanings and pronunciations. In order to introduce the characteristic that the Chinese character exists in a particular pronunciation across languages, we take the following as examples. For instance, the Chinese character "窗" (window) is pronounced as / tʂʰuaŋ/ (IPA notation) with high-level tone as / ˥/ (IPA notation) in Mandarin and / tʰaŋ ˥/ in Taiwanese. In addition to such cross-lingual variations, there are also within-lingual variations. For examples, / tʂʰuaŋ ˥/ is often mistakenly pronounced as / tsʰuaŋ ˥/ (the un-retroflex variation of / tʂʰhuaŋ ˥/) by native Taiwanese speakers. As in the case of English, which has a more complex vowel inventory than the Han language family, the words "ear" and "year" are difficult for Mandarin speakers to tell apart. In other words, pronunciation variation is in fact a natural and unavoidable phenomenon in a multi-lingual environment.

The pronunciation model plays an important role in the Chinese character-based ASR engine. It not only provides more choices during decoding if the speaker exhibits variations in pronunciation but also handles various speaking styles. As mentioned above, one Chinese character has more than two pronunciations in the combined phonetic inventory of Mandarin and Taiwanese. The factors of accent and regional migration can influence the pronunciation or speaking style of speakers too.

4. EXPERIMENTS

The experiments were conducted to validate the efficiency of the proposed one-pass recognizer. Since it is a much simpler scheme to deal with the code-switching utterances, it is worth to be adopted as long as it can achieve comparable performance as the traditional multi-pass recognition scheme.

To compare with the traditional multi-pass scheme, we have done a series of experiments on it. The procedures are shown as in figure 4. It can be divided into three stages, where the first stage is language boundary detection (LBD), the second is LID, and the final stage is the traditional mono-lingual speech recognition.

Five left-to-right routes were shown in figure 4 to connect the blocks. Each route corresponds to a series of processes in an experiment. For examples, route 0 contains three blocks, namely "Manual LBD", "Manual LID", and "Manual SR". It represents an experiment of all manual processing in the 3 stages. The other blocks in figure 4 are described in the following. The "Automatic LID" block represents the process of language identification, where the PSR as mentioned in section 2.1 was used. The "Integrated Automatic LBD & LID" represents the process of integrated language boundary detection and language identification, where the "bi-syllable probability" mentioned in section 2.2 were used. Finally, the "ASR-M" and "ASR-T" blocks represent the traditional mono-language speech recognizers for Mandarin and Taiwanese, where a normal HMM-based ASR was used. [13]

The output of route 0, labeled as R0, is the manually transcribed text, which is used as "golden text" to evaluate the accuracy of the output of the other routes, namely, from R1 to R4. The output of "Manual LID" in route 1 is the result of manual language identification, which is used as "golden LID" to evaluate the accuracy of the outputs of the other "automatic" LID subsystems, namely L2 and L3. Additionally, the bottom route represents the integrated one-pass approach proposed in this paper. The recognition results were labeled as R4, which is also compared with R0 to measure its accuracy.

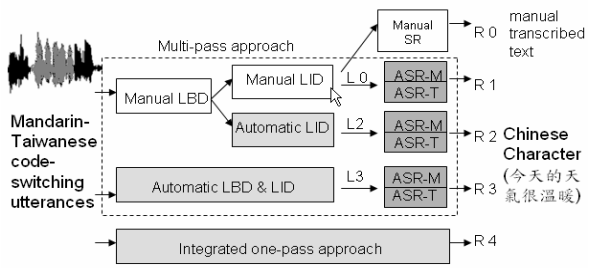

Figure 4. The experimental procedures for traditional multi-pass scheme with the proposed integrated one-pass approach in the bottom for comparison.

4.1. Mandarin-Taiwanese code-switching corpus

The speech corpus used in all the experiments were divided into two parts, namely, the training set and the testing set. The training set consists of two monolingual Taiwanese and Mandarin speech data, which includes 100 speakers. Each speaker read about 700 phonetically abundant utterances in both languages. For testing data set, another 16 speakers were asked to record 4000 Mandarin-Taiwanese code-switching utterances. Among these utterances, at least one Taiwanese phrase is embedded into a Mandarin carrier sentence. The length of each phrase is various from one to eight syllables. The statistics of the corpus used here are listed in Table 2.

	Language	No. of Speakers	No. of Words	No. of Hours
Training set	M	100	43,078	11.3
	T	100	46,086	11.2
Testing set	CS.	16	4,000	4.41

Table 2. Statistics of the bi-lingual speech corpus used for training and testing sets. M: Mandarin, T: Taiwanese, CS: code-switching utterances.

4.2. Experiment setup

The acoustic features used in ASR subsystems 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. [16] Both the language-dependent and bi-lingual HMM-based acoustic models of the syllable recognizer are trained using the corpora described above. The syllable accuracies are 63.67%, 61.69%, and 60.81% for Mandarin-only HMM, Taiwanese-only HMM and bi-lingual HMM, respectively.

In the pronunciation modeling, the average number of pronunciations for one Chinese character for each pronunciation lexicon was 1.2, 1.8 and 3.0 for Mandarin, Taiwanese and bi-lingual, respectively. Furthermore, we use the tree-structured searching nets with 2 kinds of vocabulary sizes, i.e., 10 thousands words and 20 thousands words. All the words in the testing corpus are included in the searching net and thus the out-of-vocabulary rate is zero. Additionally, the outputs of the recognizer were Chinese characters; therefore, we evaluated the performance of the ASR in terms of the Chinese character error rate (CER).

4.3. Results

The experimental results for L2 and L3, which are outcomes of the LID subsystems in figure 4, are shown in <table 3>, where the average detection rate for them are 88.05% and 76.68%, respectively.

	L2	L3
For Mandarin	89.22%	79.12%
For Taiwanese	86.87%	74.23%
Average	**88.05**%	**76.68**%

Table3. The experimental results of LID subsystems in figure 4

The experimental results for R1, R2, R3 and R4, which are the final recognition outcomes of the whole system in figure 4 are shown in <table 4>, where two kinds of vocabulary-sizes were tested in the ASR subsystems, namely, 10 thousands words and 20 thousands words. One can see that the error rates for the 20K vocabulary task are 22.59%, 28.61%, 31.76% and 20.02% for R1, R2, R3 and R4, respectively.

It is interesting that R4 outperforms R1, the latter is based on manual language segmentation and manual LID but the former used the completely automatic approach. The critical point is that R4 uses the global information and does the soft decision on LBD and LID during the decoding process. On the other hands, R1 has only local information and does the hard decision in each step before ASR, though the decision is made by human. In other words, even human could not do a correct decision under the fragment information.

	10K	20K
R1	14.22%	22.59%
R2	20.7%	28.61%
R3	23.20%	31.76%
R4	13.31%	20.02%

Table 4. The experimental results for the final recognition outcomes of the whole system in figure 4

5. CONCLUSION

In this paper, we compared two approaches to recognize a Taiwanese-Mandarin code-switching utterance as a Chinese character sequence. In the multi-pass ASR, three stages of processing are integrated, including language boundary detection, language identification and language-dependent speech recognition. We evaluated the performance of automatic approach in each stage, and also demonstrated the manual results as the golden benchmark. In each stage of the processing, the intermediate results are close to each other. On the other hand, the one-pass ASR proposed in this paper can deal with the code-switching utterance satisfactorily without extra language boundary detection and language identification. This paper presents an alternative to recognize one utterance in mixed-languages; the experimental results of performance on the Chinese character error rate demonstrate that it is a promising approach.

6. REFERENCES

[1] H. Y. Su "Code-switching between Mandarin and Taiwanese in Three Telephone Conversation: The Negotiation of Interpersonal Relationships among Bilingual Speakers in Taiwan," In Proc. of the Symposium about Language and Society, April, 2001

[2] Joyce Y. C. Chan, P. C. Ching and T. Lee, "Development of a Cantonese-English Code-mixing Speech Corpus," In Proc. of Eurospeech, 2005.

[3] C.M. Chen "Two Types of Code-Switching in Taiwan ," 15th Sociolinguistics Symposium, April 2004

[4] Joyce Y. C. Chan, P.C. Ching, Tan Lee and Helen M. Meng, "Detection of Language Boundary in Code-switching utterances by Bi-phone Probabilities," ISCSLP, 2004

[5] 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.

[6] Y. K. Muthusamy, E. Barnard and R. A. Cole "Reviewing Automatic Language Identification," IEEE Signal Processing Magazine, Vol. 11, Issue 4, 1994, pp. 33-41.

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

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

[9] J. L. Rouas, J. Farinas and F. Pellegrino, "Automatic Modeling of Rhythm and Intonation for Language Identification," in Proc. of 15th ICPhS, Barcelona, Spain, 2003, pp. 567-570.

[10]F. Cummins, "Speech Rhythm and Rhythmic Taxonomy," in Speech Prosody, Aix-en-Provence, France, 2002, pp. 121-126.

[11]J. L. Rouas, J. Farinas, F. Pellegrino and R. A. Obrecht, "Modeling Prosody for Language Identification on Read and Spontaneous Speech," In Proc. of ICASSP, 2003.

[12]H. Fujisaki, "Prosody, Information and Modeling with Emphasis on Tonal Features of Speech," in workshop on Spoken Language Processing, Mumbai, India, Joundry 2003.

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

[14]International Phonetic Association, Handbook of the International Phonetic Alphabet., Cambridge University Press, New York, NY, 1999

[15]D.C. Lyu, R.Y. Lyu, Y.C. Chiang and C.N. Hsu, "Modeling Pronunciation Cariation for Bi-Lingual Mandarin/Taiwanese SPEECH RECOGNITION," International Journal of Computational Linguistics and Chinese Language Processing, Vol. 10. no. 3. 2005, pp. 363-380

[16]D.C. Lyu, et al., "Large Vocabulary Taiwanese (Min-nan) Speech Recognition Using Tone Features and Statistical Pronunciation Modeling," In Proc. of Eurospeech, 2003