ISCA Archive
http://www.isca-speech.org/archive

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

CHINESE DIALECT IDENTIFICATION USING AN ACOUSTIC-PHONOTACTIC MODEL

Wuei-He Tsai and *Wen-Whei Chang*
Department of Communication Engineering
National Chiao Tung University, Hsinchu, Taiwan
E-mail: wwchang@cc.nctu.edu.tw

ABSTRACT

In this paper we develop hidden Markov model (HMM) based approaches to identify Chinese dialects spoken in Taiwan. This task can be aided by exploiting various characteristic features of Chinese spoken languages. The baseline system performs phonotactic analysis after the speech utterance is tokenized into a sequence of five broad phonetic classes. The sequential statistics of the resulting symbols are then used to distinguish one dialect from another. The second approach we tested is to incorporate dialect-dependent phonotactic constraints into the phonetic tokenization rather than applying these constraints after the broad phonetic classification is complete. These algorithms were evaluated using a multi-speaker speech corpus of text-independent spontaneous speech data. Simulation results indicate that the acoustic-phonotactic approach to dialect identification yields better performance with an average identification rate of 89.6%, compared to 70% for the baseline system.

1. INTRODUCION

An automatic language identification (language-ID) system takes as input speech utterances and produces as output the identity of the language being spoken. Previous work [1,2,3] on automatic language-ID suggests that sources of information useful for language discrimination include acoustic phonetics, phonotactics, prosodics, and vocabulary. Many approaches have been proposed which attempt to combine multiple information sources to distinguish one language from another. HMM-based language-ID was first proposed by House and Neuburg [3]. Their work showed that languages differ in phonotactic regularities embedded in the phonetic transcriptions of text. Although the ultimate goals and primary applications of dialect-ID are much similar as those of language-ID, porting a well-developed language-ID system to the problem of dialect-ID may present its own set of problems. This appears due to the fact that dialects are more closely related; they use the same written characters and share a common subset of phonemes.

In this paper, we propose to develop HMM-based techniques [4] for automatic identification of three major Chinese dialects spoken in Taiwan, namely, Mandarin, Holo, and Hakka. A nice feature of Chinese spoken languages is that all the characters are monosyllablic, and each syllable can be decomposed into an initial/final format. There are a total of 22 initials and 38 finals in Mandarin, 18 initials and 75 finals in Holo, and 19 initials and 65 finals in Hakka. According to the manners of articulation, each of these sub-syllables can be further classified into five broad phonetic classes (BPCs), namely, stop (A), fricative (B), affricate (C), nasal (D), and vowel or diphthong (E). Chinese dialects differ significantly from each other with respect to the frequency of occurrences of these BPCs and the order in which they occur in words. Therefore, the key to solving the problem of Chinese dialect-ID is the detection and exploitation of such phonological differences among dialects.

2. A BASELINE DIALECT-ID SYSTEM

The baseline system applied here is to perform phonotactic analysis following the broad phonetic segmentation of speech. There is a wide variance in broad phonetic patterns within and across Chinese dialects. To illustrate this, Table 1 lists all the legitimate combinations of BPC symbols used to produce an initial or final sub-syllable. An initial sub-syllable is composed of a single BPC, whereas a final sub-syllable may contain one or two BPC symbols. The baseline dialect-ID system comprises a BPC recognizer followed by a dialect-dependent phonotactic analyzer, as shown in Figure 1.

The BPC recognizer is designed to tokenize the speech utterance into a sequence of BPC symbols. Its accuracy can be aided by taking into consideration the monosyllabic nature of Chinese spoken language. As shown in Figure 2, the BPC recognizer begins with an initial/final segmentation in order to reduce the inventory size of phonologically allowed units used to produce an initial or final sub-syllable. After that, speech utterances are converted from their digital waveform representations into streams of feature vectors consisting of the lowest 10 coefficients of the mel-scaled cepstrum. The temporal structure of these feature vectors is described by using a subsyllable-based continuous HMM (CHMM) with a left-to-right topology. Each model has 9 states and its state observation probability density is modeled as a mixture of 15 underlying Gaussian densities. In the training phase, a large amount of phonetically tran-

scribed training speech is used to estimate the model parameters through the segmental *k*-means training procedure. During recognition, we employ the Viterbi decoding to find the optimal state sequence and then calculate the likelihood that the test sub-syllable is produced in each of CHMMs. Finally, the test sub-syllable is hypothesized as the BPC pattern that was used to train the maximum-likelihood model.

Using the BPC recognizer as a front-end, phonetic transcriptions of speech utterances are reduced to five-character alphabets and these samples are used to form two-state language models of each dialect. Language models designed to capture phonotactic information of each dialect can be constructed by running training speech into the BPC recognizer and computing transition probabilities between BPC symbols. In our implementation, an ergodic 2-state discrete HMM (DHMM) was used with parameters trained from the Baum-Welch re-estimation algorithm. When an unknown utterance is received, the language model takes as input the underlying BPC symbol sequence and produces as output the likelihood of the dialect being spoken. The dialect of the model with the highest likelihood is hypothesized as the dialect of the test utterance.

3. AN ACOUSTIC-PHONOTACTIC APPROACH

The baseline dialect-ID system presents its own set of problems. Firstly, it requires phonetically labeled data of each dialect for use in training the stochastic model of BPC recognizer. Secondly, observations used in language modeling will not be extracted from the cepstral feature vectors, but from a presumably correct sequence of broad-phonetic transcriptions of speech utterances. In other words, the errors due to BPC recognition would degrade the overall performance of the dialect-ID system. To compensate these shortages, we propose to incorporate dialect-specific phonotactic constraints into the broad-phonetic tokenization rather than applying these constraints after the front-end BPC recognition is complete. Figure 3 shows the block diagram of our proposed system based on an integrated acoustic-phonotactic model. It makes use of cepstral feature vectors in the dialect-ID process, thereby exploiting a larger range of phonological differences between dialects than is possible with the baseline system. The basic strategy applied here is to assume that the underlying BPC symbols of a speech utterance are produced as a probabilistic function of a 5-state Markov chain. Each of these states represents a broad phonetic class such as the stop, fricative, affricate, nasal, and vowel.

The state transition diagram of a syllable-based Markov model is shown in Figure 4. This model has a double-layer structure; it is a large ergodic HMM of which each state is built from an elementary left-to-right HMM. In our implementation, each elementary model is created using a 4-state CHMM with its observation probability density modeled as a mixture of 5 Gaussian densities per state. Observations are streams of cepstral feature vectors extracted from speech utterances, as opposed to the BPC symbol sequence used in the baseline system. Proceeding in this way, the stochastic model of language acoustics can be incorporated into the dialect-ID system without forward decoding of the underlying BPC symbol sequence. The dialect-ID system operates in two phases: training and recognition. In the training phase, an ergodic HMM is trained for every dialect to be recognized, using mel-scaled cepstral coefficients as input. A simple training approach is used to initialize the dialect-dependent model parameters, which are then refined by running the segmental *k*-means reestimation algorithm. During recognition, an unknown speech utterance is classified by first converting the digitized waveform to mel-scaled cepstral coefficients and by calculating the likelihood that these feature vectors were produced in each of the three dialects. The dialect of the model most likely to have produced the test utterance is hypothesized as the dialect of the test utterance. By allowing the system to use the phonotactic constraints during the Viterbi decoding process, the most likely dialect identified during recognition is optimal with respect to some combination of both acoustic and phonotactic information.

4. EXPERIMENTAL RESULTS

To evaluate different approaches to dialect-ID, extensive computer simulations have been conducted with various sentential utterances of different characteristics. Our effort began with the collection of a context-independent speech corpus capable of supporting dialect-ID research. It currently consists of laboratory-recorded spontaneous utterances in 3 dialects: Mandarin, Holo, and Haka. These utterances were produced by 16 speakers, including 8 males and 8 females. Two databases were used here: one for training and the other for testing. The first data set composed of 60 sentential utterances in each dialect. Each utterance has an average duration of 15 seconds. On the other hand, the speech database for use in testing consisted of 21 utterances per dialect that did not include the speech segments for training the dialect-ID system. The speech signals were digitized into 16-bit format at the sampling rate of 16 kHz. Streams of cepstral feature vectors are extracted from digitized speech utterances through the use of a mel-weighted filter-bank with the order of ten.

A preliminary experiment was first performed to examine the performance of the CHMM-based BPC recognizer. Compared with phonetically labeled data, the top-choice accuracy was measured to obtain an 81.4% recognition rate. Table 2 gives the confusion matrices for the baseline dialect-ID system based on a two-state language model. Entries along the main diagonal indicate the ratio of utterances correctly identified, while off-diagonal entries correspond to incorrect decisions. From the table we can see that speech utterances spoken in Mandarin can easily distinguish from that of Holo, but are often misjudged as Hakka. Further analysis of simu-

lation results indicates that the baseline dialect-ID system does not yield better performance with an increase in the number of states used for language modeling. This is mainly due to the fact that the phonologic pattern of Chinese dialects is predominately an alternation of initial and final sub-syllables, making it particularly suited to a two-state language model. The next step of the investigation is concerned with the accuracy of the dialect-ID system that employs an integrated acoustic-phonotactic model for information integration. Table 3 lists the corresponding dialect identification accuracy. The proposed dialect-ID system is shown to yield better performance with an average recognition score of 89.6%, compared to 70% for the baseline system. This is mainly because that for the baseline system, the approximation of the underlying phonetic transcriptions of an utterance by the broad phonetic classes decreases the phonological distinction between different dialects. Other reason for its degraded performance is due to imperfect outputs of the front-end BPC recognizer.

5. CONCLUSIONS

In this paper, we have studied the issue how to best combine acoustic and phonotactic information sources into the Chinese dialect-ID system. The basic problem with the baseline system is that its phonotactic analysis is based on the broad phonetic classes and hence cannot provide sufficient cues to make the dialect-ID decision. In addition, imperfect outputs of broad phonetic classifiers could lead to fatal errors in the subsequent phonotactic analysis. Using a hybrid HMM for information integration, the proposed method demonstrates the promise and feasibility of improving the system performance by increasing the discrimination of the language model. Because the proposed method does not require phonetically transcribed training speech, it could be easily extended to other Chinese dialects as well. Future study should be focused on determining whether such dialect-ID systems are robust against speaker's variability, or whether systems incorporating prosodic information are required to provide further improvement.

ACKNOWLEDGMENTS

This work was supported by the National Science Council, Taiwan, ROC, under Grant NSC 87-2213-E009-020.

REFERENCES

[1] Y.K. Muthusamy, E. Barnard, and R.A. Cole, "Reviewing automatic language identification," *IEEE Signal Processing Magazine*, pp. 33-41, Oct. 1994.

[2] M. A. Zissman, T. P. Gleason, D. M. Rekart, and B. L. Losiewicz. "Automatic dialect identification of extemporaneous, conversational, latin American Spanish speech," *Proceedings of the 1996 International Conference on Acoustics, Speech, and Signal Processing*, pp. 777-780, 1996.

[3] A. S. House, and E. P. Neuburg, "Toward automatic identification of the language of an utterance. I. Preliminary methodological considerations," *J. Acoust. Soc. Am.* 62, 708-713, 1997.

[4] X. D. Huang, Y. Ariki, and M.A. Jack, *Hidden Markov Models for Speech Recognition*, Edinburgh University Press, Edinburgh, 1990.

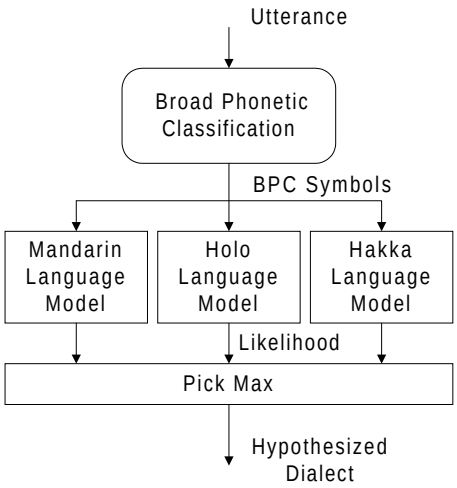

Figure 1. The baseline dialect-ID system.

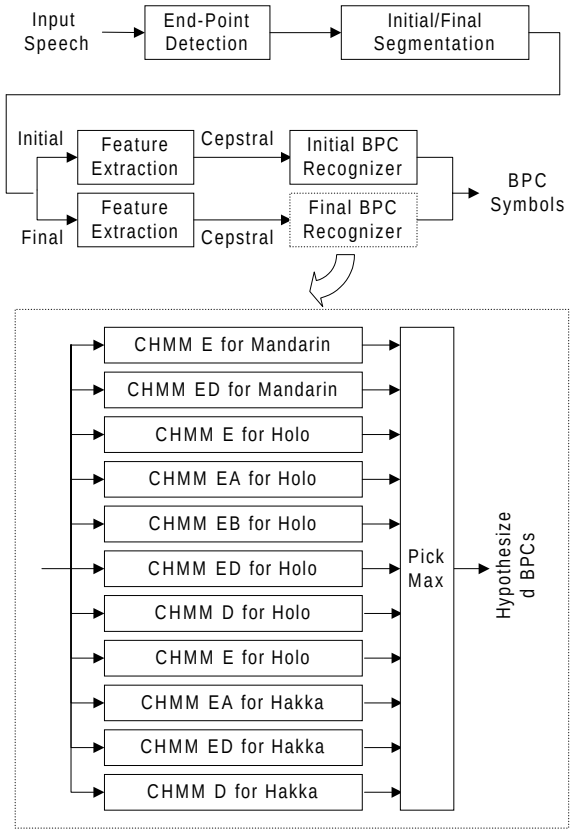

Figure 2. Broad phonetic classification.

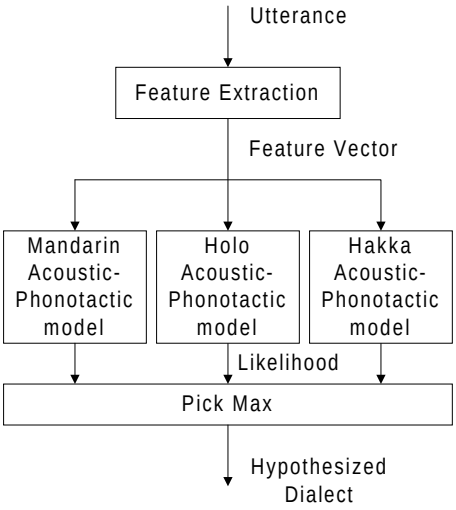

Figure 3. The dialect-ID system based on an integrated acoustic-phonotactic model.

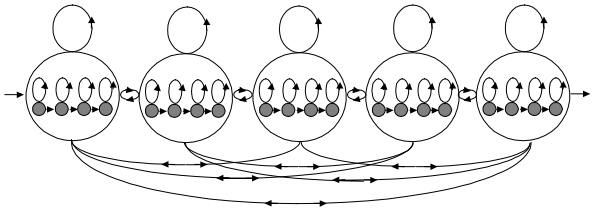

Figure 4. A hybrid HMM with double-layer structure.

Table 1: Syllable-based legitimate BPC patterns in Chinese dialects

Syllable		Dialect		
Initial	Final	Mandarin	Holo	Hakka
A	EA		√	√
B	EA		√	√
C	EA		√	√
D	EA		√	√
	EA		√	√
A	EB		√	
B	EB		√	
C	EB		√	
D	EB		√	
	EB		√	
A	ED	√	√	√
B	ED	√	√	√
C	ED	√	√	√
D	ED	√	√	√
	ED	√	√	√
A	E	√	√	√
B	E	√	√	√
C	E	√	√	√
D	E	√	√	√
	E	√	√	√
A	D		√	√
B	D		√	√
C	D		√	√
D	D		√	√
	D		√	√

Table 2: Results of dialect identification using phonotactic modeling.

Actual	Recognition		
	Mandarin	Holo	Hakka
Mandarin	**0.50**	0.00	0.50
Holo	0.00	**0.70**	0.30
Hakka	0.10	0.00	**0.90**

Table 3: Results of dialect identification using acoustic-phonotactic modeling.

Actual	Recognition		
	Mandarin	Holo	Hakka
Mandarin	**0.85**	0.09	0.06
Holo	0.11	**0.85**	0.04
Hakka	0.01	0.00	**0.99**