Automatic Identification of Chinese dialects Spoken in Taiwan

TSAI Wuei-He and CHANG Wen-Whei
Department of Communication Engineering
Chiao Tung University, Hsinchu, Taiwan

Abstract

In this paper an approach to Chinese dialect identification based on the sequential information of broad phoneme classes is described. The proposed system uses a set of category-dependent HMMs in parallel to perform broad phoneme classification, followed by phonotactic analysis using an SRN-based identifier. Furthermore, the monosyllabic property of Chinese utterances is exploited to reduce the complexity of broad phoneme classification as well as providing fine input feature patterns. It is shown that our proposed approach allows the system to differentiate three Chinese dialects from each other in a multi-speaker environment.

1 Introduction

Automatic dialect identification (dialect-ID) is a challenging problem in spoken language system. In essence, an automatic dialect-ID system takes as input speech utterances and produces as output the dialect being spoken. Although the ultimate goals and primary applications of dialect-ID are much similar as those of language-ID [1, 2], porting a well-developed language-ID system to the problem of dialect-ID may series degrade the identification performance. The main reason for this is that dialects are usually a group of more closely related languages, in the sense that they have common written form and/or share much the same structures of pronunciation.

In this paper, we propose to develop a technique which identifies three major Chinese dialects spoken in Taiwan, namely, Mandarin, Holo, and Hakka. The basic strategy applied here is to perform phonetic tokenization followed by phonotactic analysis [3]. It is aimed to exploit the phonotactic information embedded in the sequential statistics of the phonetic transcription. In the light of the basic characteristics of Chinese speech, irrespective of dialect, an isolated Chinese character is pronounced as a syllable which can be phonetically decomposed into an initial/final format and tone. Furthermore, according to the manners of articulation, each initial and final sub-syllables can be classified into five broad phoneme classes (BPCs): stop (A), fricative (B), affricate (C), nasal (D), and vowel or diphthong (E). Assume that each dialect has its own phonotactic structure that governed the combinations of different BPCs in a dialect, one can then distinguish one dialect from another by means of extracting their own phonotactic information. The architecture configuration of the dialect identifier used in this letter is shown in Fig. 1. To implement the dialect-ID system, a set of broad phoneme classifiers based on hidden Markov models (HMMs) was created to perform phonetic tokenization, followed by modeling sequential statistics of phonetic transcription through the simple recurrent network (SRN).

2 Broad Phoneme Classification

The accuracy of phoneme class partitioning can be aided by taking advantage of the monosyllabic property of Chinese utterances. Particularly, we propose to classify the phone-like units in terms of initial and final sub-syllables in order to reduce the inventory size of units of which Chinese dialects is composed. Table 1 lists all the legitimate BPC patterns observed in the three dialects. It is clear that any initial sub-syllable consists of one of the BPCs, whereas a final sub-syllable may contain one or two BPCs.

The broad phonetic classifier considered here employs a multi-dialectal BPC-recognizer to tokenize the spoken utterance into a sequence of BPC symbols. It operates in two phases: training and recognition. Prior to starting the training and recognition, the speech utterances are converted from their digital waveform representations in a stream of feature vectors consisting of ten mel-cepstral coefficients as well as their first derivatives. In the training phase, a separate left-to-right HMM [4] is created for each initial and final sub-syllables in each of the target dialects. Each HMM has 8 states and the output probability density function is modeled as a mixture of 10 underlying Gaussian densities per state. The segmental *k*-means training procedure is used for this study to estimate the model parameters. During the recognition phase, we apply the Viterbi algorithm to find the optimal state sequence and then calculate the matching score of comparing the test sub-syllable with each of the category-dependent models. Finally, the test sub-syllable is hypothesized as the phonetic pattern that was used to train the maximum likelihood model.

By tokenizing the speech waveform, the statistics of the resulting phonetic symbols can then be used to perform dialect identification. This is because that dialects differ significantly from each other with respect to the frequency of occurence of these phonetic symbols and the order in which they occur in syllables. To illustrate this, we show in Fig. 2 the statistical distribution of phonetic symbols encountered in three Chinese dialects.

3 Dialect Identification

Using the broad phoneme classifier as a front-end, the phonotactic information of the underlying dialect can be extracted by using a SRN where previous outputs of hidden nodes are delayed and then feedback to the input layer [5]. The input layer receives data from the broad phoneme classifier, the hidden layer models the phonotactic structure, and the output layer provides the hypothesis of the dialect identified. Taking the benefit of the monosyllablic property of spoken Chinese, we use syllable-level BPC pattern as the input to the network. Furthermore, in order to keep equal-distance of the input patterns, each syllable was presented to the network as a fifteen-dimensional binary vector. Table 1 also lists the associated code of all the legitimate BPC patterns to be presented in the input layer.

In the SRN, the activation function of hidden neuron j at time n is defined as

$$H_j(n) = \sum_i w_{ji} x_i(n) + \sum_l r_{jl} h_l(n-1),$$

where $x_i(n)$ is the input neuron i at time n, w_{ji} is the feedforward connection strength from input neuron i to hidden neuron j, r_{jl} is the recurrent connection strength from the delayed hidden neuron l to hidden neuron j, and

$$h_l(n\text{-}1) = f[H_l(n\text{-}1)],$$

where $f(\alpha)$=1 / (1+e$^{-\alpha}$) is a sigmoid function. The activation function of the output neuron k at time n is defined as

$$y_k(n) = f[O_k(n)],$$

and

$$O_k(n) = \sum_j W_{kj} h_j(n).$$

The network was trained using back-propagation learning algorithm based on gradient descent optimization in order to reduce the output error. Assume that the number of syllables of an utterance is L, then the output error E is defined as

$$E = \frac{1}{2}\sum_{k=1}^{K}[T_k - \frac{1}{L}\sum_{n=1}^{L} y_k(n)]^2,$$

where T_k is the output target function, which is selected as:

$$T_k = \begin{cases} 1 & \text{if } k \in \text{training neurons} \\ 0 & \text{otherwise} \end{cases}.$$

Then the weights are adjusted by using a modification of the delta rule:

$$w_{ji}(t+1) = w_{ji}(t) - \eta(t)\frac{\partial E}{\partial w_{ji}(t)},$$

$$W_{kj}(t+1) = W_{kj}(t) - \eta(t)\frac{\partial E}{\partial W_{kj}(t)},$$

$$r_{jl}(t+1) = r_{jl}(t) - \eta(t)\frac{\partial E}{\partial r_{jl}(t)},$$

where $\eta(t)$ is the learning rate at the t-th iteration. To implement this, the partial derivative terms are actual calculated by the chain rule. We iteratively adjust the weights of the SRN consisting of 15 input neurons, 5 hidden neurons, and 3 output neurons. Finally, the identification result is determined by the neuron with the largest output response.

4 Experimental Results

To test the validity of the proposed dialect identification, extensive computer simulations have been conducted with various sentential utterances of different characteristics. Two databases were used here: one for training the HMMs as well as the SRN, and the other for use in recognition. The first data set composed of 15 sentential utterances per dialect was generated by 2 male speakers. On the other hand, the speech database for use in identifying an unknown dialect consisted of 5 utterances that did not include the speech segments for training. Each utterance is, on average, 15 seconds long. The speech signals were digitized into 16-bit format at a rate of 16 kHz. According to the statistics, there were a total of 4131 initial sub-syllables and 4119 final sub-syllables providing.

A preliminary experiment was first performed to examine the validity of the HMM-based broad phoneme classifier. Compared with phonetically labeled data, correct and incorrect decisions were recorded. The top-choice accuracy was first measured to obtain a 81.4% recognition rate. Table 2 and 3 summarizes the classification results for the initial and final sub-syllable. Each entry corresponding to ith row and jth column represents the probability of classifying broad phoneme class i as broad phoneme class j. Therefore, entries along the main diagonal indicate the ratio of utterances correctly identified, while off-diagonal entries correspond to incorrect decisions. The classifier is shown to perform well since the majority of the decisions are along the main diagonal.

Next, computer simulations were conducted to examine whether Chinese spoken dialects can be accurately identified through neural network mapping. Table 4 and 5 list the experimental results for automatic identification of three Chinese dialects in content-inside and content-outside tests. We obtained an average 92.2% and 83.3% on three Chinese Dialects, respectively. From the tables we can see that the dialect pair identification rate is the lowest in the case of Mandarin and Hakka. It also indicates that the utterances spoken in Holo can be easily distinguished from that of Mandarin or Hakka.

5 Conclusions

The combined use of a HMM-based broad phoneme classifier and a SRN-based phonotactic model has been proposed for dialect identification. The initial/final structure in Chinese speech is also incorporated to further refine the dialect-ID system. Validation of the proposed system was confirmed via simulations on identification of three Chinese dialects spoken in Taiwan. It is worthwhile in future studies to extend the method towards a plurality of other Chinese dialects, as well as enhancing the robustness to the speaker's diversity.

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.

[5] S. Schreiber, D.A. Cleeremans, and J.L. McClelland, "Graded state machines: The representation of temporal contingencies in simple recurrent networks," *Machine Learning*, vol. 7, pp. 161-193, 1991.

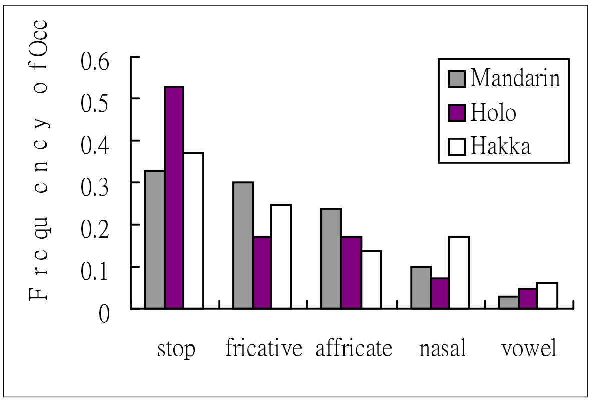

Fig. 2 Statistics of broad phoneme classes in initial sub-syllable.

Table 2 Classification results of initial sub-syllable.

Actual	Recognition				
	A	B	C	D	E
A	**0.84**	0.04	0.04	0.03	0.05
B	0.06	**0.84**	0.07	0.02	0.02
C	0.06	0.07	**0.85**	0.01	0.01
D	0.02	0.02	0.01	**0.92**	0.02
E	0.06	0.02	0.01	0.01	**0.90**

Table 3 Classification results of final sub-syllable.

Actual	Recognition				
	EA	EB	ED	E	D
EA	**0.95**	0.01	0.00	0.04	0.01
EB	0.05	**0.64**	0.07	0.23	0.01
ED	0.03	0.02	**0.84**	0.10	0.02
E	0.05	0.08	0.11	**0.74**	0.01
D	0.02	0.01	0.04	0.00	**0.93**

Table 4 Identification results of three Chinese dialects (content-inside test).

Actual	Recognition		
	Mandarin	Holo	Hakka
Mandarin	**0.87**	0.10	0.03
Holo	0.03	**0.97**	0.00
Hakka	0.07	0.00	**0.93**

Table 4 Identification results of three Chinese dialects (content-outside test).

Actual	Recognition		
	Mandarin	Holo	Hakka
Mandarin	**0.70**	0.10	0.20
Holo	0.00	**1.00**	0.00
Hakka	0.20	0.00	**0.80**

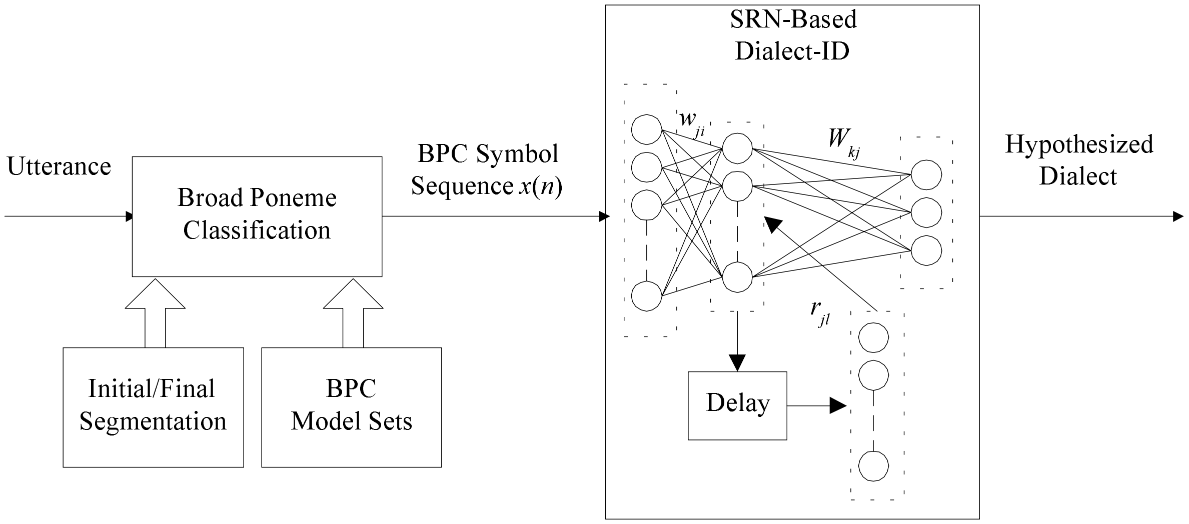

Fig.1 The architecture configuration of the dialect identifier.

Table 1 Syllable-based legitimate BPC patterns in Chinese dialects

Syllable		15 neurons	Dialect		
Initial	Final	in input layer	Mandarin	Holo	Hakka
A	EA	100000000110000		√	√
B	EA	010000000110000		√	√
C	EA	001000000110000		√	√
D	EA	000100000110000		√	√
	EA	000000000110000		√	√
A	EB	100000000101000		√	
B	EB	010000000101000		√	
C	EB	001000000101000		√	
D	EB	000100000101000		√	
	EB	000000000101000		√	
A	ED	100000000100010	√	√	√
B	ED	010000000100010	√	√	√
C	ED	001000000100010	√	√	√
D	ED	000100000100010	√	√	√
	ED	000000000100010	√	√	√
A	E	100000000100000	√	√	√
B	E	010000000100000	√	√	√
C	E	001000000100000	√	√	√
D	E	000100000100000	√	√	√
	E	000000000100000	√	√	√
A	D	100000000000010		√	√
B	D	010000000000010		√	√
C	D	001000000000010		√	√
D	D	000100000000010		√	√
	D	000000000000010		√	√