Chinese dialect identification using segmental and prosodic features

Wen-Whei Chang and Wuei-He Tsai
Department of Communications Engineering, National Chiao-Tung University, Hsinchu, Taiwan, Republic of China

(Received 9 September 1999; revised 14 April 2000; accepted 29 June 2000)

Several approaches to Chinese dialect identification based on segmental and prosodic features of speech are described in this paper. When using segmental information only, the system performs phonotactic analysis after speech utterances have been tokenized into sequences of broad phonetic classes. The second scheme comprises prosodic models which are trained to capture tone sequence information for individual dialects. Also proposed is a novel approach that examines differences between Chinese dialects at broad phonetic and prosodic levels. These algorithms were evaluated via a multispeaker read-speech mode. Simulation results indicate that the combined use of segmental and prosodic features allows the proposed system to discriminate among three major Chinese dialects spoken in Taiwan with 93.0% accuracy. © *2000 Acoustical Society of America.* [S0001-4966(00)03410-X]

PACS numbers: 43.72.Ne [DOS]

I. INTRODUCTION

Automatic language identification (language ID) is an active research area with the goal of multilingual information access. A language-ID system takes test utterances as inputs and produces the identities of the languages being spoken as outputs. Previous work[1] suggests that the language characteristics are represented in the segmental and prosodic features of speech utterances. Segmental features can be acoustic–phonetic, which refers to the acoustic realizations of phonemes, or phonotactic, which refers to the rules governing combination of various phonemes in a language. Prosodic information is encoded in the pitch, amplitude, and duration variations that span across segments. Although language-discriminating information can be found at various levels, how to best combine them for reliable language ID is still an unsolved problem. A language-ID approach based only on phonotactics was applied by House and Neuburg[2] using a hidden Markov model (HMM) trained on phonetic transcriptions of text. Recent research demonstrates that further enhancement can be realized by additionally incorporating prosodic information[3] and acoustic–phonetic information.[3–5] In using these approaches, it is assumed that the acoustic structures of individual languages can be explored by segmenting speech into basic sound units such as phonemes. Languages can then be identified by computing features within and across segments that are likely to capture the relevant phonetic and prosodic aspects of these languages.

Until recently, research in Chinese language processing was almost exclusively aimed at voice dictation of Mandarin Chinese.[6] However, hundreds of Chinese dialects exist and may be linguistically divided into seven major groups, namely, Mandarin, Holo, Hakka, Wu, Yue, Xiang, and Gan.[7] There are minor differences within each group, but there are major differences of such magnitude between groups that the groups sound mutually unintelligible. The interconnections between the Chinese dialects are in fact as complicated as those among the family of Romance languages, such as French, Spanish, and Italian. This motivated our research into trying to devise a system for determining the identities of spoken Chinese dialects. During the initial stage of developing our dialect-ID system, special efforts were made to discriminate among three major Chinese dialects spoken in Taiwan, Mandarin, Holo, and Hakka. Porting a well-developed language-ID technique to the problem of Chinese dialect ID may present its own set of problems. This is because various Chinese dialects are more closely related than the Romance languages; they use the same grammar and written characters, and include many homonyms that share the same pronunciation. System design approaches that consider Chinese language characteristics are believed to be the key to providing better solutions to the dialect-ID problem.

II. THEORY

A. Overview

The key to solving the problem of Chinese dialect ID is the detection and exploitation of characteristic features that distinguish dialects from one another. From a linguistic standpoint, the greatest differences among Chinese dialects are in the area of phonology, the least in the area of grammar. Vocabulary differences fall between these extremes. A distinctive feature of the Chinese language is that all the characters are monosyllabic. Traditional descriptions of Chinese divide syllables into combinations of *initials* and *finals* rather than into individual phonemes. An *initial* is the consonant onset (C_1) of a syllable, while a *final* consists of a nucleus (U) and an ending (C_2), where C_2 is a consonant and U can be a vowel or diphthong. There are 22 *initials* and 38 *finals* in Mandarin, 18 *initials* and 75 *finals* in Holo, and 19 *initials* and 65 *finals* in Hakka. These *initials* and *finals* can be further decomposed into more basic sound units such as phonemes and broad phonetic classes (BPCs). Table I gives a list of phonetic elements of Chinese language. De-

TABLE I. Phonetic elements used in Chinese language.

	Broad phonetic classes	Phonemes
Syllable onset (C_1)	Stop (S)	[p] [t] [k] [ph] [th] [kh] [b] [g] [l]
	Fricative (F)	[f] [h] [s] [sh] [shi] [j] [v]
	Affricate (A)	[j] [ji] [ch] [chi] [tz] [ts]
	Nasal (N)	[m] [n] [ng]
Syllable nucleus (U)	Vowel or diphthong (V)	[a] [e] [i] [o] [u] [è] [ai] [ao] [au] [ei] [eu] [ia] [ie] [io] [iu] [oa] [oe] [oi] [ou] [ua] [ue] [ui] [uo] [iao] [uai]
Syllable ending (C_2)	Stop (S)	[p] [t] [k]
	Fricative (F)	[h]
	Nasal (N)	[m] [n] [ng]

pending on the manner of articulation, phonemes can be categorized into five BPCs including the stop (S), fricative (F), affricate (A), nasal (N), and vowel or diphthong (V). Hereafter, we use this five-character alphabet in referring to BPCs. In this study, we propose using BPCs rather than phonemes as the bases for dialect discrimination. This is mainly because BPCs are relatively invariant across Chinese dialects, eliminating the need for developing a standard phonemic transcription system appropriate for all dialects. Chinese dialects differ in the frequency of BPC occurrences and the order in which BPCs occur in syllables. Table II lists eligible BPC combinations used in individual dialects. It shows that an *initial* is composed of a single BPC while a *final* generally contains one or two BPCs.

TABLE II. Eligible combinations of broad phonetic classes in Chinese syllables (S: stop, F: fricative, A: affricate, N: nasal, V: vowel or diphthong).

Syllable					
Initial	*Final*		Dialect		
Onset	Nucleus	Ending	Mandarin	Holo	Hakka
S	V	S		√	√
F	V	S		√	√
A	V	S		√	√
N	V	S		√	√
	V	S		√	√
S	V	F		√	
F	V	F		√	
A	V	F		√	
N	V	F		√	
	V	F		√	
S	V	N	√	√	√
F	V	N	√	√	√
A	V	N	√	√	√
N	V	N	√	√	√
	V	N	√	√	√
S	V		√	√	√
F	V		√	√	√
A	V		√	√	√
N	V		√	√	√
	V		√	√	√
S		N		√	√
F		N		√	√
A		N		√	√
N		N		√	√
		N		√	√

Another important feature of Chinese language is the existence of lexical tones for syllables, which means syllables may have the same phonetic compositions, but different lexical meanings when spoken with different tones. Chinese is spoken with four basic tones, traditionally labeled *ping* (level), *shaang* (rising), *chiuh* (departing), and *ruh* (entering). The most distinctive feature of *ruh*-tone syllables is that they end in one of the three stops /p/, /t/, and /k/, and thus, are much shorter in duration. The other tones consist of syllables that end in vowels, fricatives, or nasals, and differ from one another in their pitch contours. Each of these four basic tones is further split into *yin* and *yang* categories according to whether the *initial* is voiced or unvoiced, thus giving rise to an eight-tone system for spoken Chinese. Pitch contours are generally affected by *initials* in this way: pronunciation of syllables with voiced *initials* begins at a lower pitch than that used for those with unvoiced *initials*. Table III shows the eight traditional tonal categories of Chinese language. It is important to note that only a few modern dialects preserve this eight-tone system intact; in many dialects two or more of these tonal categories have been merged. For example, Mandarin does not have a *ruh* tone and requires only four tones to cover pitch variations within syllables. Chinese dialects differ not only in the number of distinct tonal categories they use, but also in the acoustic realizations of similar tones. To illustrate this, we follow Chao's system[8] for tonal notation and represent pitch height on a 5-point scale, on which 1 is low, 2 half-low, 3 middle, 4 half-high, and 5 high. Using this notation, tones can be described by indicating their beginning and ending points; in a few cases tones have concave or convex contours making it necessary to include one turning point as well. For example, in Mandarin the *shaang* tone is associated with a tone value 214, and thus falls first from the half-low point to the low point and then rises to the half-high point. Details of the tones used in individual dialects, along with their tone values on the 5-point scale, are shown in Table IV.

B. Probabilistic framework

Designing a reliable dialect-ID scheme requires that stochastic models be used to summarize some of the most relevant aspects of language acoustics. As suggested by Hazen and Zue,[3] we formulated the problem using a segment-based probabilistic framework. Choosing an appropriate representation of acoustic information is the first step in applying statistical methods to solving the dialect-ID problem. The specific types of feature measurements considered here are pitch contour and mel-cepstrum. Pitch contour is tone related, whereas the mel-cepstrum is used for determining phonetic transcriptions of utterances. Speech signals are preprocessed to extract these features for every 40-ms Hamming-windowed frame with 10-ms frame shifts.

Consider a text consisting of a succession of J Chinese characters, each of which is pronounced as a single syllable. In describing syllable j, let t_j represent the lexical tone and let $\{w_{3j-2}, w_{3j-1}, w_{3j}\}$ represent a BPC triplet used to tokenize the syllable onset, nucleus, and ending. The underlying tone and BPC sequences of an utterance can thus be defined by $T=\{t_1, t_2, \ldots, t_J\}$ and $W=\{w_1, w_2, \ldots, w_{3J}\}$, re-

TABLE III. The eight traditional tonal categories of Chinese language.

	Tonal category			
Initial class	*ping*	*shaang*	*chiuh*	*ruh*
Unvoiced	*yinping* (upper level)	*yinshaang* (upper rising)	*yinchiuh* (upper departing)	*yinruh* (upper entering)
Voiced	*yangping* (lower level)	*yangshaang* (lower rising)	*yangchiuh* (lower departing)	*yangruh* (lower entering)

spectively. Notice that not all syllables have three phonemes and hence null frames may exist in the BPC sequence W. From a modeling perspective, speech signals can be considered as templates formed by concatenating sequences of acoustic segments. Each segment roughly represents a BPC and is characterized by variable-length vector sequences of two acoustic features: pitch and mel-cepstrum. Mel-cepstral features are measured frame by frame and are of the following form:

$$\mathbf{c}=\{c_{s(1)},\ldots,c_{s(2)-1},c_{s(2)},\ldots,c_{s(3)-1},\ldots,c_{s(3J)},\ldots,c_K\}, \quad (1)$$

where K is the number of frames in an utterance and $s(l)$ denotes the starting frame for segment l. Similarly, $\mathbf{p}=\{p_1,p_2,\ldots,p_K\}$ is the sequence of K vectors representing the pitch measurements of an utterance.

The first of the three dialect-ID experiments we conducted was based on sequential BPC statistics, the second on pitch contour, and the third on a combination of pitch and segmental features. We begin by presenting the formulation of a probabilistic framework employed in the third experiment. Speech signals were first processed to extract pitch and mel-cepstral features, and these feature measurements were used to train stochastic models for every dialect to be recognized. Dialects were then identified by matching test utterances with the stochastic model of each dialect and calculating the *a posteriori* probability $\Pr(L_i|W,T,\mathbf{c},\mathbf{p})$ of the measurements $\{\mathbf{c},\mathbf{p}\}$ and the sequences $\{W,T\}$ for each dialect L_i. According to the maximum-likelihood decision rule, the classifier should decide in favor of a dialect $\hat{L}$, satisfying

$$\hat{L}=\arg\max \log \Pr(L_i|W,T,\mathbf{c},\mathbf{p}). \quad (2)$$

Since language-discriminating information can be found at various levels, it is worth evaluating the relative contribution towards dialect ID that each source of information provides. To advance with this, it is more convenient to rewrite Eq. (2) as

$$\hat{L}=\arg\max_i\{\log \Pr(\mathbf{c}|\mathbf{p},W,T,L_i)+\log \Pr(\mathbf{p},T|W,L_i) + \log \Pr(W|L_i)\}, \quad (3)$$

where the *a priori* dialect probability $\Pr(L_i)$ is assumed to be uniform and hence ignored. When only access to the BPC sequence W and mel-cepstral features $\mathbf{c}$ are available, the dialect-ID process can be simplified as follows:

$$\hat{L}=\arg\max_i[\log \Pr(\mathbf{c}|W,L_i)+\log \Pr(W|L_i)]. \quad (4)$$

A two-step approach, BPC recognition followed by phonotactic analysis, has been shown to be effective for language ID[9] and hence was employed in experiment 1. During recognition, the front-end BPC recognizer decodes the mel-cepstral features $\mathbf{c}$ into a BPC sequence $\hat{W}$ using the following expression:

$$\hat{W}=\arg\max_W \log \Pr(\mathbf{c}|W,L_i). \quad (5)$$

After that, the phonotactic analysis component calculates the likelihood of BPC sequence $\hat{W}$ being produced in each of the dialects. The most likely language model is thereby identified, and the dialect $\hat{L}$ of that model is taken as the hypothetical dialect of the test utterance in the following form:

$$\hat{L}=\arg\max_i \log \Pr(\hat{W}|L_i). \quad (6)$$

The approach employed in experiment 2 incorporated only the prosodic information as represented by pitch contour $\mathbf{p}$ and the tone sequence T. This reduced dialect-ID processing to the following:

$$\hat{L}=\arg\max_i \log \Pr(\mathbf{p},T|L_i) = \arg\max_i[\log \Pr(\mathbf{p}|T,L_i)+\log \Pr(T|L_i)]. \quad (7)$$

Implicit in this equation is the assumption that the prosodic model $\Pr(\mathbf{p},T|L_i)$ can be expressed as the product of two separate models: the pitch model $\Pr(\mathbf{p}|T,L_i)$ and the tone model $\Pr(T|L_i)$. The pitch model accounts for the various realizations of Chinese tones that may occur across dialects, and the tone model accounts for the tone statistics within each dialect.

TABLE IV. Tonal categories used in Mandarin, Holo, and Hakka, along with their tone values on the 5-point scale of pitch height. (Underlining of a tonal value indicates that the tone in question is shorter than those which are not underlined.)

Tonal category		Dialect		
		Mandarin	Holo	Hakka
ping	*yin*	55	43	44
	yang	35	21	12
shaang	*yin*	214	51	31
	yang			
chiuh	*yin*	51	35	42
	yang		212	
ruh	*yin*		43	21
	yang		21	44

III. EXPERIMENT 1: BPC RECOGNITION FOLLOWED BY LANGUAGE MODELING

A. Introduction

The proposed approach was motivated by previous experiments[2] showing that languages can be distinguished solely by means of sequential BPC statistics. Our system consisted of BPC recognizers followed by phonotactically motivated language models, and included two subsystems. The first processed utterances using a bank of dialect-dependent acoustic models in parallel, and output phonetic elements associated with the most likely model. Given the mel-cepstral features $\mathbf{c}$, the most likely BPC sequence $\hat{W}=(\hat{w}_1,\hat{w}_2,\ldots,\hat{w}_{3J})$ was determined using Eq. (5). By assuming mel-cepstral measurements are statistically independent across segments, we were able to solve for the individually most likely BPC $\hat{w}_l$ for segment l as follows:

$$\hat{w}_l=\arg\max_{w_l}\log \Pr(\mathbf{c}^{(l)}|w_l,L_i),\quad 1\leqslant l\leqslant 3J, \tag{8}$$

where $\mathbf{c}^{(l)}=\{c_{s(l)},\ldots,c_{s(l+1)-1}\}$ represents the mel-cepstral measurements for segment l. The second subsystem calculated the likelihood that BPC sequence $\hat{W}$ would be produced in each of the dialects. The maximum-likelihood classifier hypothesized that $\hat{L}$ was the dialect of the test utterance using Eq. (6).

B. Method

1. Speech corpora

Eight male speakers, aged between 18 and 30, were employed to collect Chinese speech corpora that contained utterances spoken in three dialects, Mandarin, Holo, and Hakka. For this study we attempted to avoid the speaker bias by using speakers who are fluent in the three dialects studied. The first corpus, denoted as DB-1, was designed to provide a corpus of sentential utterances for the training and testing of language models. Our text materials consisted of 30 folk-tale passages in colloquial style. Each passage consisted of 26 to 32 syllables, and the texts were grouped by passages into sets of 20 and 10. The set with 20 passages was used for training, and the set with 10 passages was used for testing. All the speakers were asked to read the text three times, once in each of the three dialects. Each syllable was spoken almost as if isolated from the adjacent syllables, but without pauses and with syllables joined in a normal-feeling and fluent manner. The number of training utterances in DB-1 was 480 (20 passages$\times$8 speakers$\times$3 dialects), and the number of test utterances in DB-1 was 240 (10 passages$\times$8 speakers $\times$3 dialects). The second corpus, denoted as DB-2, contained two sets of monosyllabic utterances produced by each of the eight speakers, one for training and one for testing in our BPC recognition experiment. Each set consisted of one reading of all eligible syllables, including 408 Mandarin syllables, 808 Holo syllables, and 708 Hakka syllables. All of the utterances in the speech corpora were recorded in a relatively quite environment, and then sampled at 16 kHz with 16-bit precision.

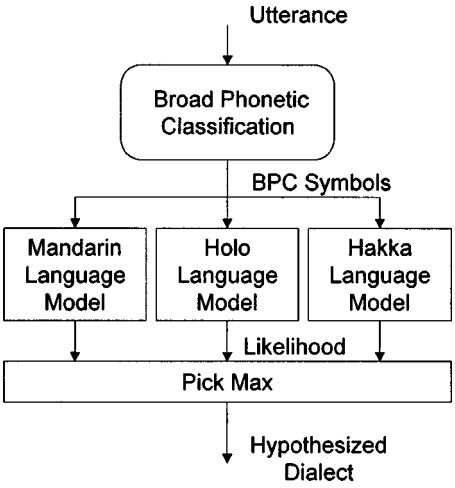

FIG. 1. Illustration of dialect-ID process which performs broad phonetic classification followed by phonotactic analysis.

2. Procedure

The basic idea was to perform phonotactic analysis after speech utterances had been tokenized into BPC sequences. A block diagram of the proposed dialect-ID system is shown in Fig. 1. In designing the BPC recognizer, we found the phonological structures of Chinese syllables could be used to advantage in broad phonetic segmentation of speech. As shown in Fig. 2, the BPC recognizer begins with *initial/final* segmentation in order to reduce the inventory size of allowable units of which Chinese syllables are composed. After

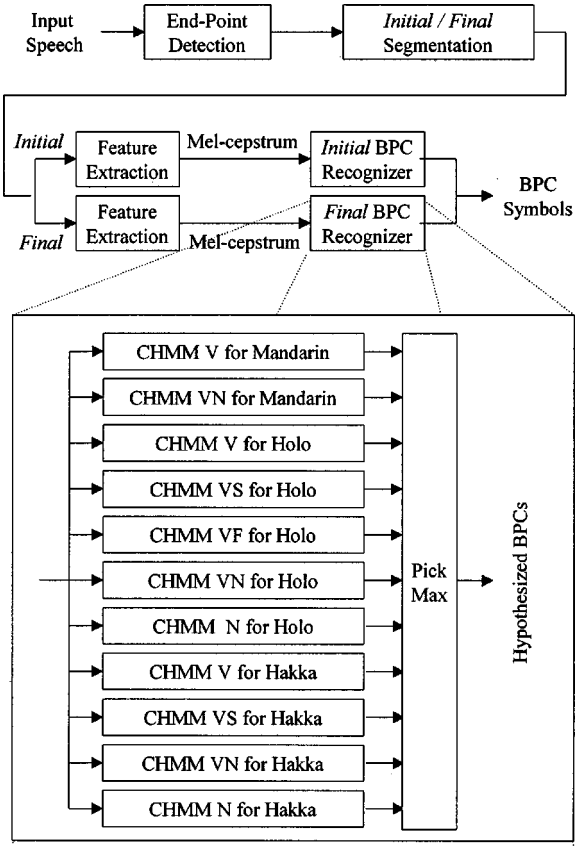

FIG. 2. Detailed description of broad phonetic classification (S: stop, F: fricative, A: affricate, N: nasal, V: vowel or diphthong).

TABLE V. Dialect-ID results based on sequential BPC statistics.

	Recognition		
Actual	Mandarin	Holo	Hakka
Mandarin	**0.50**	0.00	0.50
Holo	0.00	**0.80**	0.20
Hakka	0.10	0.00	**0.90**

that, speech utterances are converted from their digital waveform representations into streams of feature vectors consisting of the lowest ten coefficients of the mel-cepstrum. The temporal structures of these feature vectors are described using a segment-based continuous density hidden Markov model (CHMM) with a left-to-right topology. Each model has nine states and its state observation probability density is modeled as a mixture of 15 underlying Gaussian densities. The training set of the DB-2 corpus was used to estimate the model parameters according to the segmental *k*-means algorithm.[10] During recognition, we employ Viterbi decoding to find the optimal state sequence associated with observed acoustic features and then calculate the likelihood that test subsyllables were produced in each of the CHMMs. Finally, test subsyllables are tokenized as BPC patterns 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 language models that perform the dialect-ID task. Training and test utterances were chosen from the DB-1 corpus. In the training phase, a separate language model was created for each dialect by running the training utterances into the BPC recognizer and computing transition probabilities between successive BPCs. In our implementation, an ergodic five-state discrete observation HMM (DHMM) was used with parameters trained according to the Baum–Welch reestimation algorithm.[11] When an unknown utterance is received, the language model receives as its input the sequence of recognized BPCs and produces as output the likelihood of the dialects being spoken. Finally, the dialect of the language model that predicts the utterance with the highest likelihood is taken as the dialect of the test utterance.

C. Results and discussion

We first examined the performance of the CHMM-based BPC recognizer used as the front-end processor in this experiment. The top-choice accuracy achieved a recognition score of 81.4%, as compared with phonetically labeled data. Further analysis of our results showed that recognition errors occurred mainly in classification of fricative sounds as stops. Sequential statistics on these BPCs were then input to the language models that performed dialect ID. Table V shows a confusion matrix using the identification results for this experiment. The rows of the confusion matrix correspond to the dialects actually being spoken and the columns indicate the dialects identified. The classification accuracy of 73.3% indicates that BPC sequence information alone cannot provide sufficient cues for making accurate dialect-ID decisions. As the table shows, dialect-ID results have a bias towards identifying test utterances as Hakka. We speculate that this might be attributed to the difference in closeness between pairs of the three dialects. Support for such a speculation can be found in Cheng,[12] in which the correlation method was applied to phonological elements to quantify affinity among Chinese dialects. It was found that the highest degree of closeness among the three dialects studied was in the Mandarin–Hakka pair, the middle in the Holo–Hakka pair, and the least in the Mandarin–Holo pair. To elaborate further, we follow the method in Juang and Rabiner[13] and compute the probabilistic distance for measuring the dissimilarity between pairs of DHMM-based language models. Considering two HMMs specified by the parameter sets λ_1 and λ_2, the probabilistic distance measure is defined by

$$\mathcal{D}_s(\lambda_1,\lambda_2)=\frac{1}{2M}\left[\log\frac{\Pr(\mathcal{O}^{(2)}|\lambda_1)}{\Pr(\mathcal{O}^{(2)}|\lambda_2)}+\log\frac{\Pr(\mathcal{O}^{(1)}|\lambda_2)}{\Pr(\mathcal{O}^{(1)}|\lambda_1)}\right], \tag{9}$$

where $\mathcal{O}^{(j)}$ is the sequence of M observations generated by the model λ_j. It was found that as the distance measure increased, there was a corresponding improvement in dialect-ID accuracy. Our results showing the measured distances for pairwise dialect-ID tasks are plotted in Fig. 3. It is clear from this figure that the dialect-ID system using BPC sequence information was more successful when applied to distinguishing Mandarin from Holo. We also found that the system did not yield better performance with an increase in the number of states used in the language model, perhaps because Chinese-language phonologic pattern is predominantly an alternation of C_1, U, and C_2, making it particularly suited to a three-state language model.

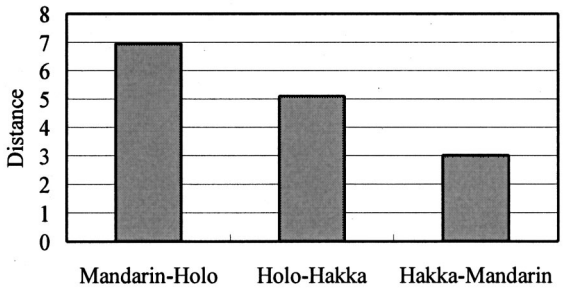

FIG. 3. The measured distances between language models for pairwise dialect ID.

IV. EXPERIMENT 2: PROSODIC MODELING

A. Introduction

Most current approaches to language ID make little or no use of prosodic measures, despite evidence showing that prosodic information is useful in human identification of languages. The main reason for this is the difficulty of finding an appropriate feature set that captures linguistically relevant prosodic information. Early approaches that incorporated a limited number of prosodic features did not produce satisfactory results.[14] Improved language-ID performance requires better modeling of prosody using a large set of feature measurements such as those proposed by Thyme-Gobbel and Hutchins.[15] They used a total of 220 features including av-

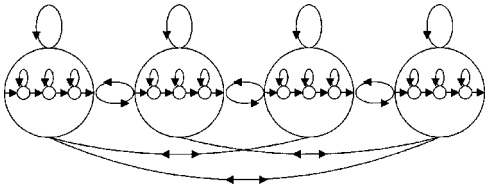

FIG. 4. Illustration of CHMM-based prosodic model used for Mandarin.

erages, deltas, standard deviations, and correlations of measures deduced from pitch contours, syllable durations, amplitudes, and rhythms. However, increasing the amount of detail in the models can be computationally costly. This problem can be alleviated in the dialect-ID task, mainly because the Chinese tonal system provides an efficient way to describe pitch contour dynamics. The goal of this experiment was to investigate whether Chinese dialects can be identified using only pitch measurements **p** and the tone sequence *T*. Accordingly, we sought to determine the most likely dialect $\hat{L}$ using Eq. (7).

B. Method

The utterances used to train and test the prosodic models were taken from the DB-1 corpus. Prosodic models designed to capture the tonal statistics of individual dialects were created using a composite HMM structure. The underlying hypothesis is that the tones in utterances are produced as probabilistic functions of ergodic Markov chains. Each state corresponds to one of the tonal categories and is built from an elementary three-state left-to-right HMM. Because at most, one turning point exists in standard pitch contours, it suffices to postulate that the pitch contour dynamics corresponding to various tones can be modeled by three states. Figure 4 shows the state transition diagram required to represent the Mandarin prosodic model. In our implementation, each elementary model was created using a CHMM with the observation probability density modeled as a mixture of five Gaussian densities per state. Observations were independent streams of pitch and differential pitch extracted from digitized speech using the simple inverse filter tracking (SIFT) algorithm.[16] In order to describe pitch contour dynamics, CHMM-based prosodic models for each of the target dialects were constructed using parameters trained with the segmental *k*-means algorithm. When a test utterance is received, the prosodic model takes the pitch measurements as input and produces the likelihood of the dialects being spoken as output. The dialect of the model most likely to have produced the test utterance observations is taken as the dialect of the test utterance.

C. Results and discussion

Table VI shows a confusion matrix using dialect-ID results for experiment 2. We can see in the table that pitch information alone allowed the system to identify three dialects with an accuracy score of 64.3%, indicating that prosodic features are highly useful in Chinese dialect ID. The success of this prosodic model arises from its ability to exploit not only differences among dialects that exist in tonal statistics, but also differences in the realizations of similar tones in those dialects. To elaborate further, we plotted the probabilistic distances that measured dissimilarities between pairs of CHMM-based prosodic models. The results, shown in Fig. 5, indicate that from a prosodic standpoint, differentiation between Mandarin and Hakka is easier than other pairwise dialect-ID tasks. While these results suggest that pitch-based prosodic features possess language-discriminating information, it may be overly ambitious to hope that a single prosodic model can be designed to capture all of the complexities of a dialect. Therefore, a system that considers segmental and prosodic information might be required to achieve a higher degree of dialect-ID accuracy.

TABLE VI. Dialect-ID performance of the pitch-based prosodic model.

	Recognition		
Actual	Mandarin	Holo	Hakka
Mandarin	**0.73**	0.12	0.15
Holo	0.30	**0.54**	0.16
Hakka	0.11	0.23	**0.66**

V. EXPERIMENT 3: SYSTEM INTEGRATION

A. Introduction

The approach employed in experiment 1, BPC recognition followed by language modeling, has been shown to be effective for language ID,[9,17] and may be considered as representing the state of the art. The main disadvantage of this approach is that the observations used in language modeling are not extracted directly from mel-cepstral coefficients, but rather from the imperfect outputs of the front-end BPC recognizer. To compensate for this shortcoming, we propose incorporating dialect-specific phonotactic constraints into the phonetic tokenization rather than applying these constraints after BPC recognition has been completed. Furthermore, we believe that prosody may provide many benefits as an enhancement to the state-of-the-art technique by acting as a secondary source of language-discriminating information. Although a similar approach to combining segmental features with prosodic modeling was presented by Hazen and Zue,[3] the basic design here is quite different due to the special characteristics of the Chinese language. It has long been recognized that there was no explicit model to describe prosodic information, especially for nontonal languages such as English. Therefore, Hazen and Zue[3] proposed a prosodic

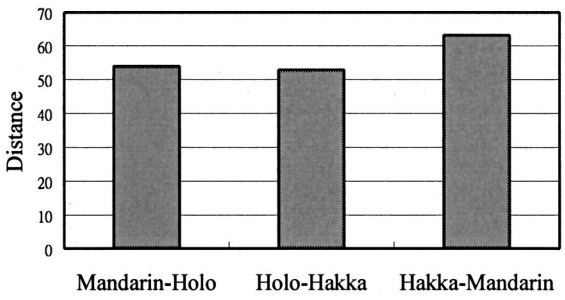

FIG. 5. The measured distances between prosodic models for pairwise dialect ID.

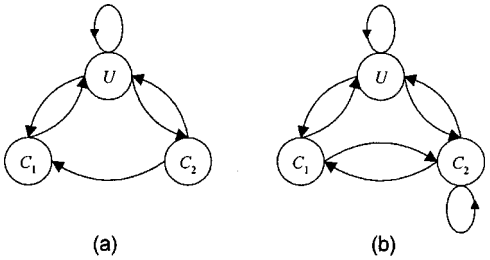

FIG. 6. State transition diagrams of composite models used for (a) Mandarin and (b) Holo and Hakka (C_1: syllable onset, U: syllable nucleus, C_2: syllable ending).

model that only captures simple statistical information about the fundamental frequency and segmental duration of an utterance. However, the prosodic modeling problem is not a serious obstacle in our dialect-ID system, mainly because Chinese-language prosodic units are annotated in the tone levels. For this experiment we attempted to determine the most likely dialect using Eq. (2), by combining segmental and prosodic information within a unified framework.

B. Method

The basic idea was to combine segmental and prosodic information to achieve a higher degree of dialect-ID accuracy. The first step toward realization was to use a three-state Markov chain to model the changing statistical characteristics present in BPC production. Each state corresponds to one acoustic segment, such as syllable onset (C_1), syllable nucleus (U), or syllable ending (C_2). The choice of model topology depended on the syllable types phonologically allowed in the target dialect. Holo and Hakka allow six syllable types, represented by U, C_1U, UC_2, C_1UC_2, C_1C_2, and C_2, while Mandarin allows only four syllable types, U, C_1U, UC_2, and C_1UC_2. Figure 6 shows the state transition diagrams required to model the individual dialects. This model has a composite structure; it is a large Markov chain in which each state is built from a bank of elementary left-to-right HMMs. Note that the final state of one elementary HMM is connected to the initial state of the following elementary HMM by a null transition.

In order to integrate prosodic features into the composite model, we propose using tonal BPCs rather than BPCs as the bases for dialect discrimination. Every tonal BPC can, in fact, be considered as a combination of two components: one of the possible tones plus one of the five possible BPCs. Since pitch is defined only for voiced speech, the pertinent tone-related portions of syllables are the nuclei from which distinctive pitch changes are perceived. Recognizing this, we needed only to superimpose distinct tonal notations onto the BPC equivalents of the syllable nuclei in order to obtain a set of BPC variants that shares the same BPCs but has distinct lexical tones. To illustrate this, details of the elementary HMMs required to model the Holo syllables are shown in Fig. 7, with V_j denoting the BPC symbol V associated with the tone j. In our implementation, each elementary model uses a three-state left-to-right CHMM with its observation probability density modeled as a mixture of five Gaussian densities per state. Observations are independent streams of

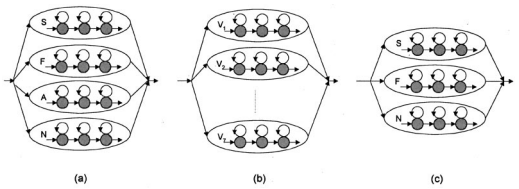

FIG. 7. Elementary HMMs required for modeling Holo (a) syllable onset, (b) syllable nucleus, and (c) syllable ending (S: stop, F: fricative, A: affricate, N: nasal, V: vowel or diphthong).

mel-cepstral coefficients and pitch measurements derived from digitized speech. By using mel-cepstral coefficients as parts of state observations, the stochastic acoustic–phonetic model can be directly incorporated into the dialect-ID process without forward decoding of underlying BPC sequences. This eliminates many of the errors made by the front-end BPC recognizer.

The dialect-ID system operates in two phases: training and recognition. The utterances used here were taken from the DB-1 corpus. In the training phase, a composite model was trained for every dialect to be recognized by running the segmental k-means algorithm. During recognition, the test utterance was classified by extracting feature vectors from digitized waveform and then 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 was taken as the dialect of the test utterance. By allowing the system to use a composite model during the Viterbi decoding process, the most likely dialect was optimal with respect to some combination of segmental and prosodic information.

C. Results and discussion

Table VII lists a confusion matrix showing the dialect-ID results for the system that examined differences between dialects at broad-phonetic and prosodic levels. In it we see that compared with the first two experiments, the effectiveness of using an integrated segmental-prosodic model for dialect ID is clearly demonstrated. The top-choice accuracy was measured to obtain a recognition score of 93.0%. Among the reasons for this success, we find that the integration of acoustic–phonetic and phonotactic information by means of a composite HMM model increases phonological discrimination across dialects. It is also important to note that the language-discriminating power of the system can be improved by using tonal BPCs as the bases for its elementary HMMs.

TABLE VII. Dialect-ID performance of the integrated segmental-prosodic model.

	Recognition		
Actual	Mandarin	Holo	Hakka
Mandarin	**0.94**	0.03	0.03
Holo	0.00	**0.86**	0.14
Hakka	0.00	0.01	**0.99**

VI. SUMMARY

This paper presents several approaches that employ varying degrees of linguistic traits discriminate among three major Chinese dialects spoken in Taiwan. The first approach includes two subsystems, the first of which uses acoustic models to recognize broad-phonetic classes, and the second uses language models to identify target dialects. Simulation results indicate that while segmental features are useful in Chinese dialect ID, other sources of information are likely to help in distinguishing between dialects with greater accuracy. The importance of incorporating prosodic information is reflected in the observation that using only pitch contour dynamics allows the system to identify three dialects with 64.3% accuracy. Recognizing this, we studied the issue of how to best combine segmental and prosodic features within one system. Using a composite HMM for information integration, our proposed method demonstrates the promise of improving system performance by increasing its language-discriminating power. The main attraction of the proposed dialect-ID approach arises from its being tailored specifically to the Chinese language, and the ease with which it can be extended to identify other Chinese dialects as well.

Although fairly good performances were reported in these experiments, more work is needed to further validate the proposed dialect-ID system for a wider range of speech corpora. Specifically, it should include combined male and female speech. It should also include continuous speech that preserves tone sandhi[8] during recording even when the utterances are being stretched temporally for easy processing. When dealing with combined male and female speech, a promising approach is to first determine the speaker's gender and then to perform dialect identification using the models of selected gender. Automatic gender classification has been previously investigated using the difference of position of the first and second formants between male and female speakers.[18] In extending the current system for continuous speech, more sophisticated tone models are necessary because tone patterns of syllables in continuous speech are subject to various modifications by sandhi rules. In our view, this problem should be approached with the same modeling techniques as were used in tone recognition of continuous speech. Previous work[6] on Mandarin speech suggests that it suffices to describe the tone sandhi and coarticulation effects using a total of 23 context-dependent tone models. However, it is still unclear at this time how many such models will be required for accurate modeling the tone variations across syllables in Holo and Hakka.

ACKNOWLEDGMENTS

This research was supported by the National Science Council, Taiwan, ROC, under Grant No. NSC87-2213-E009-039. The authors are very grateful to the unknown reviewers and the associate editor, Dr. Douglas O'Shaughnessy, for their careful readings of this paper and their constructive suggestions.

[1] Y. K. Muthusamy, E. Barnard, and R. A. Cole, ''Reviewing automatic language identification,'' IEEE Signal Process. Mag. **4**, 33–41 (1994).
[2] 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 (1977).
[3] T. J. Hazen and V. W. Zue, ''Segment-based automatic language identification,'' J. Acoust. Soc. Am. **101**, 2323–2331 (1997).
[4] M. A. Zissman, ''Comparison of four approaches to automatic language identification of telephone speech,'' IEEE Trans. Speech Audio Process. **SAP-4**, 31–44 (1996).
[5] L. F. Lamel and J. L. Gauvain, ''Language identification using phone-based acoustic likelihoods,'' in *Proceedings of the 1994 International Conference on Acoustics, Speech, and Signal Processing* (IEEE, Piscataway, NJ, 1994), pp. 293–296.
[6] L. S. Lee, ''Voice dictation of Mandarin Chinese,'' IEEE Signal Process. Mag. **14**, 63–101 (1997).
[7] S. R. Ramsey, *The Languages of China* (Princeton University Press, Princeton, NJ, 1987).
[8] Y. R. Chao, *A Grammar of Spoken Chinese* (University of California, Berkeley, CA, 1968).
[9] Y. K. Muthusamy, K. Berkling, T. Arai, R. A. Cole, and E. Barnard, ''A comparison of approaches to automatic language identification using telephone speech,'' in *Proceedings of 3rd European Conference on Speech Communication and Technology* (European Speech Communication Association, Grenoble, France, 1993), pp. 1307–1310.
[10] L. R. Rabiner, J. G. Wilpon, and B. H. Juang, ''A segmental *k*-means algorithm training procedure for connected word recognition,'' AT&T Tech. J. **65**, 21–32 (1986).
[11] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, ''A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains,'' Ann. Math. Stat. **41**, 164–171 (1970).
[12] C. C. Cheng, ''Quantifying affinity among Chinese dialects,'' J. Chin. Linguist. **3**, 78–112 (1991).
[13] B. H. Juang and L. R. Rabiner, ''A probabilistic distance measure for hidden Markov models,'' AT&T Tech. J. **64**, 391–408 (1985).
[14] Y. K. Muthusamy, ''Segmental approach to automatic language identification,'' Ph.D. thesis, Oregon Graduate Institute of Science & Technology (1993).
[15] A. Thyme-Gobbel and S. E. Hutchins, ''On using prosodic cues in automatic language identification,'' in *Proceedings of the 1996 International Conference on Spoken Language Processing* (Philadelphia, PA, 1996), pp. 1768–1771.
[16] J. D. Markel, ''The SIFT algorithm for fundamental frequency estimation,'' IEEE Trans. Audio Electroacoust. **AU-20**, 367–377 (1972).
[17] T. J. Hazen and V. W. Zue, ''Automatic language identification using a segment based approach,'' in *Proceedings of the 3rd European Conference on Speech Communication and Technology* (European Speech Communication Assoc., Grenoble, France, 1993), pp. 1303–1306.
[18] R. Vergin, A. Farhat, and D. O'Shaughnessy, ''Robust gender-dependent acoustic–phonetic modelling in continuous speech recognition based on a new automatic male/female classification,'' in *Proceedings of the 1996 International Conference on Spoken Language Processing* (Philadelphia, PA, 1996), pp. 1081–1084.