INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY 7, 129–140, 2004

An Innovative Prosody Modeling Method for Chinese Speech Recognition

GANG PENG AND WILLIAM S.-Y. WANG
Language Engineering Laboratory, Department of Electronic Engineering, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong
gpeng@ee.cityu.edu.hk
eewsyw@cityu.edu.hk

Abstract. This paper presents an innovative method for prosody modeling in Chinese speech recognition. Our method first evaluated the reliability of the prosodic information by which the recognition system dynamically tunes the balance between the spectral scores and prosodic scores. The basic idea of this method is to use prosodic knowledge based on its reliability. The higher the reliability, the more the prosodic information contributes to recognition. Thus, this method will not introduce extra errors but will incorporate more knowledge into the recognition system. Experimental results showed that this method reduced the relative word error rate by as much as 52.9% and 46.0% for Mandarin and Cantonese digit string recognition tasks, respectively. When incorporating tone information into Cantonese Large Vocabulary Continuous Speech Recognition (LVCSR) via the proposed method, a 20.16% relative character error rate reduction was obtained.

Keywords: Chinese dialects, speech recognition, prosody modeling, context-dependent

1. Introduction

Prosody is a collection of supra-segmental features, notably intensity, duration and F_0 (Fundamental Frequency), that are critical to human speech perception. The prosodic information of speech is closely related to various linguistic and non-linguistic features, such as word meaning, syntactic structure and the speaker's emotion and intention. In speech communication, it plays a very important role in the transmission of information. From the applied point of view, many prosody modeling methods (Burshtein, 1996; Ferguson, 1980; Huang and Seide, 2000; Lau et al., 2000; Lee et al., 1990; Levinson, 1986; Rabiner, 1984a, b, 1989; Ramesh and Wilpon, 1992; Russell and Moore, 1985; Wilpon et al., 1991) have been proposed to incorporate prosodic knowledge into speech recognition systems.

Although prosodic information has been proved useful in speech recognition, many Automatic Speech Recognition (ASR) systems process only spectral cues. They ignore or deliberately remove prosodic cues, due to the large variations in such cues and the lack of efficient algorithms that can extract duration and fundamental frequency from speech signals with high accuracy. Especially in noisy environments, the prosodic information extracted from speech signals may be very unreliable. Nevertheless, extending speech recognition systems to human performance levels will require exploiting all available cues, including prosodic information.

This paper will address an innovative prosody modeling method in speech recognition for Mandarin and Cantonese. We concentrate on a practical approach to tone and duration modeling within the framework of Hidden Markov Models (HMMs).

In the next section, phonological and phonetic properties of both Mandarin and Cantonese will be described. Then, the innovative prosodic modeling method, Reliability Guided Prosody Modeling (RGPM), will be introduced in Section 3. The databases used for evaluation and experimental results will be described in Section 4. Finally, conclusions and future work will be discussed in Section 5.

Tone			
[Initial]	Final		
	[Medial]	Vowel	[Ending]

Figure 1. Syllable structure of Chinese dialects. [ ] means optional (Wang, 1973)

2. Chinese Phonology and Phonetics

Each Mandarin and Cantonese utterance can be viewed as a concatenation of mono-syllabic sounds. Typically, each Chinese character is pronounced as a monosyllable. However, a Chinese character may have multiple syllable pronunciations, called polyphones, while one pronunciation can be shared by several characters, in which case they are called homophones.

Chinese is a tonal language in which F_0 patterns are used to build words, much as consonants and vowels do, while in English, the F_0 patterns of a syllable indicate stresses and intonations. There are 1761 so-called tonal syllables in Cantonese and 1471 tonal syllables in Mandarin; ignoring tonal differences, there are about 625 and 420 so-called base syllables, in Cantonese and Mandarin respectively.

2.1. Syllabic Structure

As shown in Fig. 1, each syllable of both Mandarin and Cantonese can be divided into an optional Initial and an obligatory Final. The Initial is a consonant, while the Final consists of a vowel complex (including a single vowel) with or without an ending, which may be a nasal, a retroflex or a stop. The retroflex does not occur in Cantonese, while the stop ending is not used in Mandarin. Furthermore, in Mandarin, there are three widely used medials, /i/, /u/ and /y/. On the other hand, in Cantonese, there is only one medial, /u/, and this medial can only occur after the velar Initials. In order to avoid this kind of skewed distribution, most phonologists view this medial, /u/, as a part of the labialized velar Initials. So phonologically, we can say there is no medial, but phonetically, there is still one medial in Cantonese.

2.2. Lexical Tones

Cantonese has a rich inventory of tones. Traditionally, it is said there are nine tones in Cantonese, as shown in Fig. 2. Each tone on checked syllables corresponds to a tone on unchecked syllables in terms of the F_0 pattern, so it also can be said there are six tones in Cantonese. In Mandarin, there are only four lexical tones, as show in Fig. 3, and one neutral tone whose F_0 contour completely depends on its immediately preceding tone, which is a highly context-dependent tone.

2.3. Initials

Table 1 lists the Initials for both Mandarin and Cantonese. The 19 Cantonese Initials are labeled with JyutPing and the International Phonetic Alphabet (IPA) (LSHK, 2002), while the 21[1] Mandarin Initials are labeled with Pinyin and IPA (Lin and Wang, 1992).

2.4. Finals

Table 2 lists the Finals for Mandarin and Cantonese. There are 53 Finals in Cantonese and 37[2] Finals in Mandarin.

3. Reliability Guided Prosody Modeling

Although prosodic information is important for speech recognition, such information extracted from speech signals may not be very reliable. If the wrong

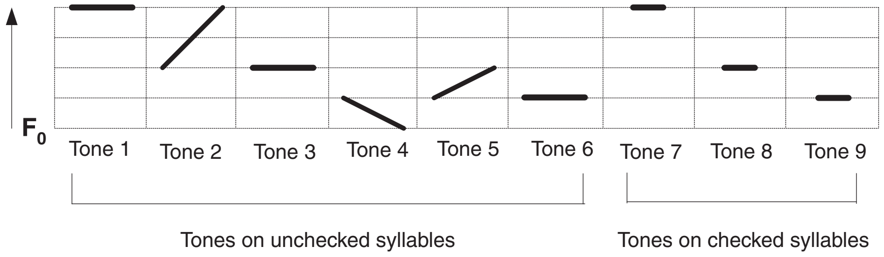


Figure 2. Lexical tones of Cantonese.

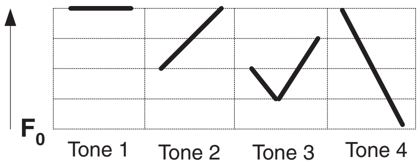


Figure 3. Lexical tones of Mandarin.

information is added to the recognizer, it may introduce extra errors. Nevertheless, the overall performance will be improved by incorporating more information into ASR. Suppose we can measure the reliability of the prosodic cues. The higher the reliability, the more the contribution from such cues will be when incorporated into the recognition system. In this way, the performance of ASR systems should be significantly improved by using such a RGPM method. Note that intensity is the easiest parameter to extract from the speech signal, and we need not estimate its reliability.

3.1. Reliability Estimation of Tone Contours

The F_0 values extracted automatically from the speech signals may not be reliable enough to allow correct recognition of the tones, due, at least in parts to the extraction algorithms, tonal coarticulation, and the quality of the speech itself.

3.1.1. Factors Related to Tone Variation. Tone variation according to different tonal contexts, i.e., tone coarticulation, has been found in the continuous speech of many tone languages, such as Mandarin and Thai (Potisuk et al., 1999; Shen, 1990; Wu, 1984; Xu, 1994, 1997). Wu (1984) reported that in trisyllabic words in Mandarin, the F_0 contours of tones vary in different tonal contexts. Shen (1990) found that tones in trisyllabic sequences in Mandarin were affected by both carryover and anticipatory[3] coarticulation. Xu (1994, 1997) studied contextual tonal variation in Mandarin systematically and indicated that the F_0 contour of a tone can sometimes be distorted beyond easy recognition without taking into consideration the preceding tones. He also pointed out that the carryover effect is larger than the anticipatory effect.

As for the effects of speaking rate on tones, Gandour et al. (1999) investigated the effects of speaking rate on Thai tones. They found that Thai tones with substantial F_0 movement (falling, high, rising) exhibit overall flatter slopes at fast speaking rates; those tones with less

Table 1. Initials of Mandarin and Cantonese.

Cantonese				Mandarin			
JyutPing	IPA	Manner	Place	Pinyin	IPA	Manner	Place
b	/p/	Plosive, unaspirated	Labial	b	/p/	Plosive, unaspirated	Labial
d	/t/	Plosive, unaspirated	Alveolar	d	/t/	Plosive, unaspirated	Alveolar
g	/k/	Plosive, unaspirated	Velar	g	/k/	Plosive, unaspirated	Velar
p	/pʻ/	Plosive, aspirated	Labial	p	/pʻ/	Plosive, aspirated	Labial
t	/tʻ/	Plosive, aspirated	Alveolar	t	/tʻ/	Plosive, aspirated	Alveolar
k	/kʻ/	Plosive, aspirated	Velar	k	/kʻ/	Plosive, aspirated	Velar
z	/ts/. /tʃ/	Affricate, unaspirated	Alveolar	z	/ts/	Affricate, unaspirated	Alveolar
c	/tsʻ/. /tʃʻ/	Affricate, aspirated	Alveolar	c	/tsʻ/	Affricate, aspirated	Alveolar
s	/s/, /ʃ/	Fricative	Alveolar	s	/s/	Fricative	Alveolar
f	/f/	Fricative	Labial-dental	f	/f/	Fricative	Labial-dental
h	/h/	Fricative	Glottal	h	/x/	Fricative	Velar
l	/l/	Liquid	Lateral	l	/l/	Lateral	Alveolar
m	/m/	Nasal	Labial	m	/m/	Nasal	Labial
n	/n/	Nasal	Alveolar	n	/n/	Nasal	Alveolar
gw	/kʷ/	Plosive, unaspirated	Labial-velar	zh	/tʂ/	Affricate, unaspirated	Retroflex
kw	/kʷʻ/	Plosive, aspirated	Labial-velar	ch	/tʂʻ/	Affricate, aspirated	Retroflex
ng	/ŋ/	Nasal	Velar	sh	/ʂ/	Fricative	Retroflex
j	/j/	Glide	Alveolar	j	/tɕ/	Affricate, unaspirated	Palatal
w	/w/	Glide	Labial	q	/tɕʻ/	Affricate, aspirated	Palatal
				x	/ɕ/	Fricative	Palatal
				r	/r/	Approximant	Retroflex

Table 2. Finals of Mandarin and Cantonese.

	Cantonese	Mandarin
Vowel	i /i/, u /u/, yu /y/, e /ɛ/, oe /œ/, o /ɔ/, aa /a/	i /i/ /ɿ/ /ʅ/, u /u/, ü /y/, a /a/, o /o/, e /ɤ/, er /ɚ/
Vowel-nasal	im /im/, in /in/, ing /ɪŋ/, yun /yn/, un /un/, ung /ʊŋ/, eng /ɛŋ/, eon /ɵn/, oeng /œŋ/, on /ɔn/, ong /ɔŋ/, am /ɐm/, an /ɐn/, ang /ɐŋ/, aam /am/, aan /an/, aang /aŋ/	an /an/, en /ən/, ang /ɑŋ/, eng /əŋ/, ian /iɛn/, in /in/, iang /iɑŋ/, ing /iŋ/, uan /uan/, uen /un/, uang /uɑŋ/, ong /uŋ/, üan /yɛn/, ün /yn/, iong /yŋ/
Diphthong	ui /ui/, ei /ei/, eoi /ɵy/, oi /ɔi/, ai /ɐi/, aai /ai/, iu /iu/, ou /ɔu/, au /ɐu/, aau /au/	ia /ia/, ie /iɛ/, iao /iɑu/, iou /iou/, ua /ua/ uo /uo/, uai /uæi/, uei /uei/, üe /yɛ/, ai /æi/, ei /ei/, ao /ɑu/, ou /ou/
Vowel-stop	ip /ip/, it /it/, ik /ɪk/, yut /yt/, ut /ut/, uk /ʊk/, ek /ɛk/, eot /ɵt/, oek /œk/, ot /ɔt/, ok /ɔk/, ap /ɐp/, at /ɐt/, ak /ɐk/, aap /ap/, aat /at/, aak /ak/	
Syllabic nasal	m /m̩/, ng /ŋ̍/	

F_0 movement (mid, low) display steeper slopes. Wu (1984) also indicated that tones in trisyllabic words can have their F_0 contours distorted at fast speaking rates.

From prior studies and our observation, the faster the speaking rate, the heavier the coarticulatory effect. We hypothesize that (1) if the speaking rate is no faster than average, then the tone coarticulatory effect will not affect the recognition of tones; (2) if the tone has no tonal context, then there is no need to consider the coarticulatory effect; and (3) if the tone has only right context or left context, then the coarticulatory effect should be weaker than when it has both contexts. Then, the effect of speaking rate on the reliability of the F_0 contours is defined as

$$f_1 = \begin{cases} 1 & \text{if } \alpha \geq 1, \text{ or without tonal context,} \\ \min\{1, 1.2\alpha\} & \text{with either left or right tonal context,} \\ \alpha & \text{otherwise,} \end{cases} \tag{1}$$

where the number '1.2' (and also the numbers in the following formula) in Eq. (1) was determined empirically, and α is the speaking rate of an utterance, defined as

$$\frac{1}{N}\sum_{i=1}^{N}\frac{d_i}{\mu_i}, \tag{2}$$

where N is the number of syllables in that utterance, d_i is the duration of the ith syllable, and μ_i is the average duration of the ith syllable in a corresponding position (final position or non-final position).

3.1.2. Factors Related to Speech Quality. When producing Mandarin Tone 3, the phonation type of the middle speech segment always changes from normal voice to fry voice[4] (Kong, 2001). It is very difficult to use traditional methods to describe the F_0 values in those segments whose voice type is fry. Technically, the result of this phenomenon is identical to situations in which the F_0 values cannot be extracted from voiced speech segments.

3.1.3. Factors Related to F_0 Extraction Algorithms. Some F_0 values may be wrongly extracted from the speech signal for several reasons. Talkin (1995) summarized the difficulties of F_0 extraction. In Mandarin and Cantonese, the following features make F_0 difficult to estimate:

(1) Sub-harmonics of F_0 often appear that are sub-multiples of the "true" F_0;
(2) Vocal tract resonances and transmission channel filtering can emphasize higher harmonics than the first harmonic, which misleads the identification by an integral multiple of F_0;
(3) In some cases, when strong sub-harmonics are present, the most reasonable objective F_0 estimate is clearly in conflict with auditory perception;
(4) Occasionally, F_0 actually jumps up or down by an octave; and
(5) In many cases, irregular voicing is present at the voice onset and offset, and very low energy may be present at the ends of syllables, which decreases wave-shape similarity in adjacent periods.

Although hundreds of F_0 extraction algorithms (Hess, 1983; Talkin, 1995) have been proposed to overcome the above difficulties, no algorithm yet can guarantee the perfect extraction of F_0. The following effects related to F_0 extraction were considered in the calculation of the reliability of the F_0 contour.

(1) As shown in Fig. 4, some parts of the F_0 contour were missing, that is, the values of F_0 were

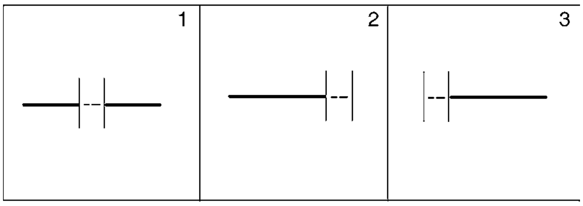


Figure 4. Some parts of the F_0 contour are missing.

not determined by the F_0 extraction algorithms. If there were missing parts in the middle of the F_0 contour, as shown in pane 1, we just discarded this information because the reliability of this incomplete F_0 was very low. If the missing parts were at the beginning or end of the F_0 contour, as shown in panes 2 and 3, then we calculated the effective rates which indicates the percentage of the whole F_0 that was determined.

$$f_2 = \begin{cases} 0 & \text{with middle missing parts,} \\ \dfrac{D_{F_0}}{D_{\text{voiced}}} & \text{otherwise,} \end{cases} \tag{3}$$

where D_{F_0} was the duration of the F_0 contour, and D_{voiced} was the duration of the voiced part of the corresponding syllable.

(2) The degree of jitter of F_0 contours. Normally, there are some jitters of the F_0 contours during the vibration of the vocal cords. Due to physiological constraints, the jitters should have some limits. Here, we used a cubic curve fitting to fit the F_0 contour, then calculated the Average Square Deviation (*ASD*, i.e., the fitting error), which was defined as

$$ASD = \frac{1}{N}\sum_{i=1}^{N}(F(i) - f(i))^2, \tag{4}$$

where N was the number of F_0 values, $F(i)$ $1 \leq i \leq N$ were the values of the F_0 contour, and $f(i)$ $1 \leq i \leq N$ were the corresponding cubic curve-fitted values of the F_0 contour.

$$f_3 = \begin{cases} 1 & \text{if } ASD \leq 10, \\ \dfrac{10}{ASD} & \text{otherwise.} \end{cases} \tag{5}$$

3.1.4. Reliability of F_0 Information. We combined the above measurements simply by multiplication. The result is the reliability of F_0 information as defined by

$$R_{F_0} = f_1 \times f_2 \times f_3, \tag{6}$$

where R_{F_0} stands for the reliability of an F_0 contour.

3.2. Reliability Estimation of Duration Information

Each Mandarin or Cantonese syllable, corresponding to a Chinese character, consists of an optional Initial and an obligatory Final. Each word is built from one or several syllables (by morphology). Each sentence is composed by concatenating several words (by syntax). As shown in Table 3, the usable duration cues for a syllable embedded in a sentence include: (1) the duration of the whole syllable (here we considered the ratio of its duration over its average duration), which is identical to the summation of the duration of its Initial and that of its Final; (2) the ratio of the duration of its Final over the duration of its Initial if the syllable had an Initial (hereafter we call it the Final-Initial duration ratio); and (3) the ratios of its duration over that of its preceding syllables in the same sentence (hereafter we call them the inter-syllabic duration ratios).

In order to make the inter-syllabic duration ratio independent of the syllables themselves, the duration of syllables embedded in a sentence was normalized by their average duration (corresponding to their intrinsic duration), with consideration of other factors such as different positions. The 2nd feature in Table 3 then was the weighted mean of the preceding inter-syllabic ratios of these normalized durations, with larger weights being given to those ratios calculated from relatively 'reliable' syllabic durations.

Due to the phonetic properties of different speech units (Initials and Finals) and the syllabic structure of Mandarin and Cantonese, the boundaries of some speech units can be identified more accurately than those of others. By studying the phonetic properties and the confusion matrix produced by the baseline

Table 3. Features used for duration modeling.

1. The Final-Initial duration ratio if the syllable being processed has an Initial.	1 feature
2. The weighted average of inter-syllabic duration ratios.	1 feature
3. The ratio (relative duration) of the duration of the syllable being processed over its average duration in corresponding position.	1 feature

recognizer, we divided the Initials (and also syllables) into two sets: a reliable duration set and an unreliable duration set. For example, the duration of fricative and affricate Initials can be identified more easily and accurately than that of other Initials.

Then, the following considerations were be taken into account when evaluating the reliability of duration information:

(1) Factors related to Initial duration, defined as:

$$d_1 = \begin{cases} 1 & \text{if the duration of Initials is reliable,} \\ 0 & \text{if duration information of Initials} \\ & \quad \text{is not available or reliable; and} \end{cases} \tag{7}$$

(2) Factors related to inter-syllabic duration ratios defined as (refer to the second duration feature):

$$d_2 = \begin{cases} 1 & \text{if some of the preceding syllables} \\ & \quad \text{have reliable duration,} \\ 0 & \text{if there are no preceding syllables,} \\ 0.5 & \text{otherwise.} \end{cases} \tag{8}$$

Finally, the reliability of duration information of the syllable or combination being processed was defined as:

$$R_D = \begin{cases} 1 & d_1 = 1 \,\&\, d_2 = 1, \\ 0.8 & \text{either } d_1 = 1 \text{ } or \text{ } d_2 = 1, \\ 0.3 & d_1 = 0 \,\&\, d_2 = 0, \\ 0.5 & \text{otherwise.} \end{cases} \tag{9}$$

3.3. Prosody Modeling Method

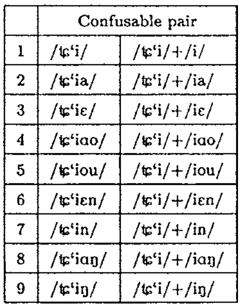
3.3.1. Tone Modeling Method. Both the F_0 values (in Hz) and the energy values (in dB) were extracted by Praat (Boersma and Weenink, 2001). The features used in our tone recognition scheme are listed in Table 4. Ten features were extracted from the syllable being processed; the remaining features were used to consider the tone coarticulatory effects. Then, Support Vector Machines (SVMs) (Burges, 1998) were used to construct the tone classifiers.

3.3.2. Duration Modeling Method. In Chinese speech recognition, some syllables or combinations are systematically related to insertion or deletion errors.

Table 4. Features for tone classification.

1. Duration of the F_0 contour of the syllable being processed; the average F_0 values and the slopes of the three uniformly divided linearly-fitted F_0 sub-contours; the means of the three corresponding log-energy sub-contours.	10 features
2. The same three features (i.e., log-energy, F_0 mean and slope) of the last sub-segment of the preceding F_0 contour and the corresponding log-energy sub-contour, and the first sub-segment of the following F_0 contour and the corresponding log-energy sub-contour.	6 features
3. Log-energies and duration of unvoiced/silent segments both before and after the F_0 contour of the syllable being processed.	4 features

Table 5. Confusable pairs related to syllable /ʨʻi/.



	Confusable pair	
1	/ʨʻi/	/ʨʻi/+/i/
2	/ʨʻia/	/ʨʻi/+/ia/
3	/ʨʻiɛ/	/ʨʻi/+/iɛ/
4	/ʨʻiɑo/	/ʨʻi/+/iɑo/
5	/ʨʻiou/	/ʨʻi/+/iou/
6	/ʨʻiɛn/	/ʨʻi/+/iɛn/
7	/ʨʻin/	/ʨʻi/+/in/
8	/ʨʻiɑŋ/	/ʨʻi/+/iɑŋ/
9	/ʨʻiŋ/	/ʨʻi/+/iŋ/

We divided the easily confusable syllables and combinations into various sets. For instance, there were nine confusable pairs related to the syllable /ʨʻi/ in Mandarin, as shown in Table 5.

The duration features selected for duration modeling are shown in Table 3. Four pairs of multivariate Gaussian functions were be used to describe the duration information of each confusable pair according to four different positions: one pair of Gaussian functions was for the starting position, one pair was for the non-final position, one pair was for the final position, and the remaining pair was for single-syllable sentence and its counterpart.

3.4. Reliability Guided Prosody Modeling Method

The reliability was further transformed to a weight by

$$w = e^{-\alpha(1-R)}, \tag{10}$$

where α was the growth rate of the exponential function, and R was the reliability of F_0, or the duration information.

Table 6. Training and test data for Mandarin and Cantonese, where '#' stands for 'Number'.

	Mandarin		Cantonese	
Properties	Training data	Test data	Training data	Test data
# of Speakers	75	17	40	12
# of Syllables	24,075	5,452	81,963	20,498
# of Sentences	6,000	1,360	22,387	5,598

Thus the final scores were obtained by combining the prosodic scores with the spectral scores produced by HMMs via the above weight. Then, the final recognition was be based on these final scores.

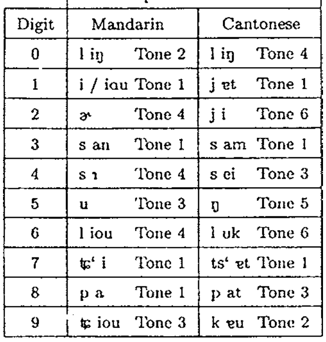
4. Preliminary Evaluation

4.1. Small Vocabulary Task

4.1.1. Database. The Mandarin database used in this study was from the Beijing Institute of Automation (Zhang et al., 1999), while the Cantonese speech database, CUDIGIT, was developed by the Chinese University of Hong Kong (Lee et al., 2002a). Table 6 gives a summary of the training and test data for our study.

4.1.2. Acoustic Modeling. Phonetically, there are five Initials and nine Finals in Mandarin digits, while there are six Initials and nine Finals in Cantonese digits, as shown in Table 7, so a total of 14 and 15 base phone models were trained for Mandarin and Cantonese, respectively. For the base phone models of both dialects, 3-state left-to-right continuous gaussian density HMMs without skipping states were adopted.

Table 7. Ten digits of Mandarin and Cantonese. In Mandarin, digit '1' has two pronunciations, and Tone 3 always changes to Tone 2 when its following tone is also a Tone 3 due to the tone-sandhi rule (Wang and Li, 1967).



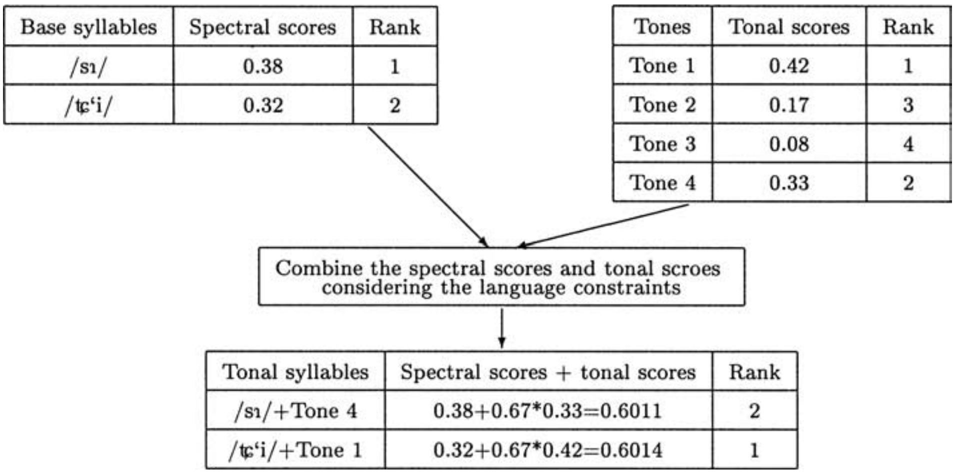
	Transcription and Tone Class	
Digit	Mandarin	Cantonese
0	l iŋ Tone 2	l iŋ Tone 4
1	i / iou Tone 1	j ɐt Tone 1
2	ɚ Tone 4	j i Tone 6
3	s an Tone 1	s am Tone 1
4	s ɿ Tone 4	s ei Tone 3
5	u Tone 3	ŋ Tone 5
6	l iou Tone 4	l ʊk Tone 6
7	tɕ‘ i Tone 1	ts‘ ɐt Tone 1
8	p a Tone 1	p at Tone 3
9	tɕ iou Tone 3	k ɐu Tone 2

Furthermore, the cross word (each digit is also a word) triphone unit has been shown to be more accurate than the whole word unit and the word internal triphone unit. A decision tree-based clustering method was used to obtain the triphone sets. This method required a phonetic question set to cluster the k_{th} HMM state of all triphones that shared the same base phone. By studying an English phonetic question set used in the ARPA Resource Management task, phonetic question sets for Mandarin and Cantonese were designed separately. Each question set asked a question like "does a triphone's left/right context belong to the specified set of this question?". Questions were designed so that all phones that appeared at the context part of a question had similar manner or place of articulation.

4.1.3. Prosody Incorporation. In Mandarin digit string recognition, the tone information is very useful for reducing the number of substitution errors. As shown in Fig. 5, suppose there is a confusable set, {/sɿ/ and /tɕ‘i/}, whose spectral scores (probability) are 0.38 and 0.32 (wrongly recognized as digit '4' only using the spectral score), respectively. The tone recognition scores for Tones 1–4 are 0.42, 0.17, 0.08 and 0.33, respectively. The reliability of the F_0 contour is 0.8 (the weight will be $e^{-2*(1-0.8)} = 0.67$). Note that these numbers are provided purely for illustration. With the language constraints of Mandarin digit strings, the syllable /tɕ‘i/ can only be associated with Tone 1, while syllable /sɿ/ can only be associated with Tone 4. Finally, the most likely result of this example would be for syllable /tɕ‘i/, with Tone 1, to be correctly recognized as digit '7' ($0.38 + 0.67 \times 0.33 = 0.6011 < 0.6014 = 0.32 + 0.67 \times 0.42$).

As for confusable patterns related to insertion and deletion errors, the process is illustrated as follows. When a syllable or combination in a confusable pair occurred, we decided whether it was a syllable or combination by comparing the weighted sum of the spectral score and the duration score. Finally, the experimental results for Mandarin are shown in Table 8.

In Cantonese digit string recognition, the tone information is not very useful for reducing the number of substitution errors, but the duration is very useful for reducing the numbers of insertion and deletion errors. The experimental results for Cantonese are shown in Table 9.



Base syllables	Spectral scores	Rank
/sɿ/	0.38	1
/tɕʻi/	0.32	2

Tones	Tonal scores	Rank
Tone 1	0.42	1
Tone 2	0.17	3
Tone 3	0.08	4
Tone 4	0.33	2

Combine the spectral scores and tonal scroes
considering the language constraints

Tonal syllables	Spectral scores + tonal scores	Rank
/sɿ/+Tone 4	0.38+0.67*0.33=0.6011	2
/tɕʻi/+Tone 1	0.32+0.67*0.42=0.6014	1

Figure 5. Schematic process of combining spectral scores with tonal scores. With the language constraints of Mandarin digit string, the syllable /tɕʻi/ can only be associated with Tone 1, while syllable /sɿ/ can only be associated with Tone 4. Without consideration of language constraints, we would need to choose among 2 × 4 = 8 options, instead of just 2 options.

4.1.4. Discussion. The above methods post-process only the 1-best recognition result of the HMM recognizers. In Mandarin, several pairs of digits are severely confused, e.g., digit '7' and digit '4', digit '8' and digit '2', which contribute more than half of the confusion errors (Peng, 2002). As shown in Fig. 5, such errors can be significantly reduced by using tone information. In Mandarin, digits '1' (for pronunciation /i/) and '5' consist of only single vowels, and digit '2' also has a heavily rhotacized vowel. These three digits, due to their short duration and vowel-only structure, are strongly coarticulated with adjacent digits in continuously spoken digit strings. Around 85% of insertion and deletion errors, e.g., the combination '7' + '1' is highly confused with '7', are related to them. In Cantonese, the digit '5' is pronounced as a syllabic nasal, which is often confused with the nasal coda of digits '0' and '3', e.g., combination '3' + '5' is confused with '3'. The above experimental results show that the duration information can be used to reduce such insertion and deletion errors.

Table 8. Word accuracy (%) of the baseline system, and baseline system with RGPM for Mandarin, where 'Acc.' stands for accuracy, 'Del.' for deletion errors, 'Sub.' for substitution errors and 'Ins.' for insertion errors. (Note that each state of HMMs finally has 8 mixtures.)

System	Word Acc.	Del.	Sub.	Ins.
Baseline	97.58%	47	40	45
Baseline + RGPM	98.86%	18	15	29

Table 9. Word accuracy (%) of the baseline system, and baseline system with RGPM for Cantonese.

System	Word Acc.	Del.	Sub.	Ins.
Baseline	98.19%	265	28	78
Baseline + RGPM	99.02%	126	28	47

The prosodic cues were used only to deal with some specific pairs, which are confused at the articulatory level, but can be distinguished at the prosodic level. The highest recognition accuracy was obtained by incorporating prosodic information into HMM-based recognizers via the above methods for both Mandarin and Cantonese digit string recognition (Lee et al., 2002a; Zhang et al., 1999). However, the usability of this idea for LVCSR is questionable if evaluated only for digits. But in the following text, this idea will be extended to incorporate tone information into Cantonese LVCSR.

4.2. *Large Vocabulary Task*

4.2.1. The Baseline LVCSR System. The acoustic models consisted of context-dependent Initial-Final models, in which each Initial model had three emitting states, while a Final model had either three or five emitting states, depending on its articulatory composition. Each emitting state consisted of eight Gaussian mixtures. The acoustic feature vector had a total of 39 components, including 12 Mel-Frequency Cepstral

Table 10. Training and test data of CUSENT, where '#', 'F' and 'M' stand for 'Number', 'Female' and 'Male' respectively. The populations of the training data and test data are completely exclusive.

Properties	Training data	Test data
# of Speakers	68(34 F, 34 M)	12(6 F, 6 M)
# of Syllables	215,604	11,677
# of Sentences	20,378	1,198

Coefficients (MFCCs), energy, their first-order derivatives and second-order derivatives. The HMMs were trained with the CUSENT database (Lee et al., 2002a), which is shown in Table 10. A decision tree-based clustering method was used to facilitate sharing of model parameters. A base syllable recognition accuracy of 79.08% was obtained for the test set from the CUSENT database.

The language model, character-based trigrams, was been built with 3,927 character entries, covering 99.99% of the Cantonese training text corpus. The training text contains about 150 million Chinese characters from WiseNews (2001). Using the test data of the CUSENT database and the above language model, a character recognition accuracy of 80.90% was obtained without using tone information.

4.2.2. Tone Recognition. The features used in the Cantonese tone recognition scheme were slightly different from those shown in Table 4. The feature *slope* was deleted, and the average F_0 (F_0 mean) was replaced with two F_0 values, one of which is at the 1/3 time point, while the other is at the 2/3 time point, of the linearly-fitted F_0 sub-contour. We made these changes because in Mandarin, the *slope* feature is distinct for different tones. But in Cantonese, the height of the F_0 contour becomes much more important because there are several level tones. In this case the selection of tone features depends on the structure of the tone system.

To facilitate the tone feature extraction, forced alignment was applied to the training and test utterances of the CUSENT database to obtain Initial-Final segmentation. Then a normalization method similar to Moving Window Normalization(MWN) was used for tone normalization (Lee et al., 2002b). In our method, the normalization window extended to the previous 0.5 second and the following 1 second of the syllable being processed, not the two preceding syllables and four succeeding syllables as used in MWN.

Then, the Support Vector Machines were used to construct the tone classifiers. An accuracy of 63.1% for tone recognition was obtained.

4.2.3. Tone Incorporation During Lattice Generation. The basic idea was to add tonal contribution during syllable lattice generation. When a syllable in one path reaches its end state, it may be added to the syllable lattice. If it is added to syllable lattice, then it will be expanded to six possible tonal syllables (illegitimate tonal syllables will be omitted). The diagram of this expansion is shown in Fig. 6. Determining how to distribute the total tonal contribution to each tonal syllable is the crucial point of the tonal syllable lattice generation. Because only voiced frames have F_0 values, we defined the total tonal contribution as

$$S_{\text{Tone}} = C_1 \times L, \tag{11}$$

where C_1 was a language-related constant (30 was selected empirically for Cantonese), and L was the number of voiced frames of the syllable being processed. If S_{Tone} is equally distributed to each tonal syllable, no tonal contribution is added to the tonal syllable lattice generation, because an equal amount of tonal score will be added to each path of the lattice. However, each tone may be assigned a different recognition score by the tone classifier based on the tone-related feature vector (tone token). These recognition scores are then sorted in a descending order. Thus the first score corresponds to the best candidate; the second score corresponds to the second best candidate, and so on. Then S_{Tone} will be distributed to each tonal syllable in proportion to its recognition score.

As shown in Table 11, the performance of the LVCSR system improved from 80.90% to 84.62% when tone information was incorporated via the proposed method. Furthermore, the effect of the reliability of tonal contours on tone recognition was studied. Figure 7 shows the relationship between reliability and tone recognition accuracy. The reliability information was incorporated in the following way: if the reliability was larger than a threshold θ (0.5 was selected empirically for θ), then some portion of the lower tone recognition score was moved to its immediately preceding better one according to the ratio, $(R_{F0} - \theta)/C_2$, where C_2 was a constant (12 was selected empirically for C_2), and R_{F0} was the reliability estimation of the F_0 contour. Otherwise, some portion of the higher score was moved to its immediately following one according to the ratio, as much as $(\theta - R_{F0})/C_2$, and we

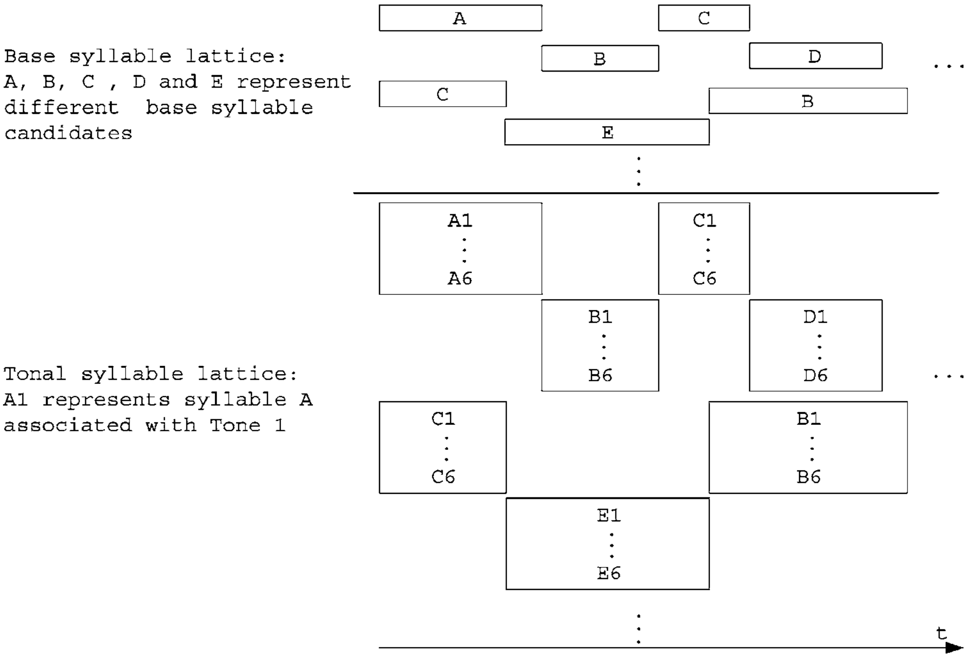


Figure 6. Base syllable lattice expansion with tone information.

should keep the same order (the originally higher score was still higher after the adjustment). Finally, an accuracy of 84.75% was obtained. Table 11 gives detailed evaluation results for individual test speakers. For seven of the twelve speakers, the incorporation of reliability information led to further improvement in the overall performance of the LVCSR system. Only one speaker encountered performance degradation. This

Table 11. Performance of integrated system.

		Character accuracy				
Speaker	Base syllable accuracy (%)	Without tone (%)	With tone (%)	Improvement (%)	With tone and reliability info. (%)	Further improvement (%)
01 (F)	75.81	83.12	86.16	3.04	86.84	0.68
02 (F)	83.66	85.50	88.80	3.30	89.00	0.10
03 (F)	73.52	79.97	80.73	0.76	80.92	0.19
04 (F)	78.00	79.98	85.50	5.52	85.50	0.00
05 (F)	82.33	76.88	83.85	6.97	83.96	0.11
06 (F)	77.78	78.13	83.33	5.20	83.65	0.32
07 (M)	83.86	80.23	83.28	3.05	83.28	0.00
08 (M)	84.00	84.59	87.38	2.79	87.38	0.00
09 (M)	76.89	79.53	82.84	3.31	82.74	−0.10
10 (M)	79.57	81.49	84.96	3.47	84.96	0.00
11 (M)	75.52	80.39	83.45	3.06	83.55	0.10
12 (M)	78.24	82.35	85.17	2.82	85.40	0.23
Overall	79.08	80.90	84.62	3.72	84.75	0.13

M: male; F: female

Table 12. Comparison with other systems.

		Character accuracy			
	Base syllable accuracy (%)	Without tone (%)	With tone (%)	Improvement (%)	Relative error rate reduction (%)
CUHK system (Lee et al., 2002b)	75.69	75.43	76.61	1.18	4.80
Our system	79.08	80.90	84.75	3.85	20.16

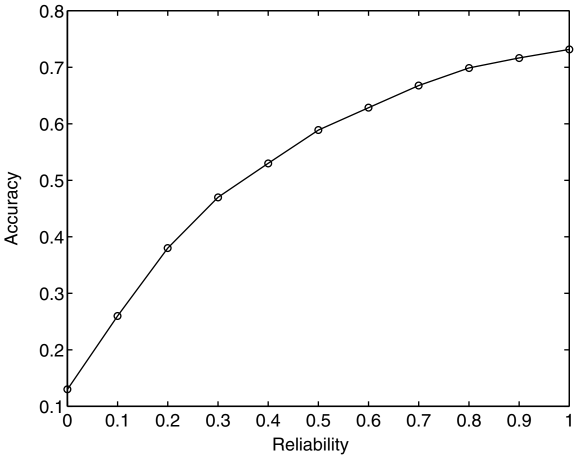


Figure 7. Accuracy versus reliability.

confirmed that the hypothesis that RGPM is useful for LVCSR.

4.2.4. Comparison with Other Systems. When the tone information was incorporated into the Cantonese LVCSR system by the proposed method, the recognition performance improved from 80.9% to 84.75%. Table 12 compares our system and the system reported in Lee et al. (2002b). The proposed method significantly improved the recognition performance.

5. Conclusions and Future Work

A new method for prosodic modeling called RGPM has been presented, in which the reliability of prosodic cues are first evaluated; the higher the reliability, the more the prosodic cues contribute to the final decision. This method produces 52.9% and 46.0% relative word error rate reduction evaluated on digit string recognition tasks for Mandarin and Cantonese, respectively.

When applied to Cantonese LVCSR, RGPM was shown to be effective. RGPM would be a promising prosody modeling method for large vocabulary Chinese speech recognition, especially in noisy environments where the prosodic information becomes less reliable.

Acknowledgment

The work described in this paper was supported by a grant from City University of Hong Kong (Project No. 7001327).

We are grateful to the anonymous reviewers of this work for their careful reading and for many thoughtful suggestions and comments.

Notes

1. The null Initial of /i/-start Finals is usually labeled as /j/, while the null Initial of /u/-start Finals is usually labeled as /w/, so it can be said there are 23 Initials in Mandarin.
2. In Mandarin, the Final /ɿ/ is associated exclusively with Initials /ts/, /tsʻ/ and /s/, and /ʅ/ is associated exclusively with Initials /tʂ/, /tʂʻ/ and /ʂ/. In the Pinyin system, the above two Finals and the Final /i/ are labeled with the same symbol ‘i’, so there are only 35 Finals in the Pinyin system of Mandarin.
3. The carryover effect is the coarticulation effect of the preceding tone on the following tone, while the anticipatory effect is the coarticulation effect of the following tone on the preceding tone.
4. During the period of fry voice, the F_0 values are extremely low; the pitch periods become very irregular; and some noise will appear.

References

Boersma, P. and Weenink, D. (2001). Praat: Doing phonetics by computer [Online]. Available: http://www.fon.hum. uva.nl/praat/

Burges, C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*. Boston: Kluwer Academic Publishers, vol. 2, pp. 121–167.

Burshtein, D. (1996). Robust parametric modeling of durations in Hidden Markov Models. *IEEE Transactions on Speech and Audio Processing*, *4*(3):240–242.

Ferguson, J.D. (1980). Variable duration models for speech. *Proceedings of Symposia on the Application of Hidden Markov Models to Text and Speech*. New-Jersey: Princeton, pp. 143–179.

Gandour, J., Tumtavitikul, A., and Satthamnuwong, N. (1999). Effects of speaking rate on Thai tones. *Phonetica*, *56*:123–134.

Hess, W. (1983). *Pitch Determination of Speech Signals: Algorithms and Devices*. Berlin: Springer-Verlag.

Huang, Hank C.-H. and Seide, F. (2000) Pitch tracking and tone features for Mandarin speech recognition. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, pp. 1523–1526.

Kong, J.-P. (2001). Study on dynamic glottis through high-speed digital imaging. Ph.D. thesis, City University of Hong Kong.

Lau, W., Lee, T., Wong, Y.W., and Ching, P.C. (2000). Incorporating tone information into Cantonese large-vocabulary continuous speech recognition. *Proceedings of the 2000 International Conference on Spoken Language Processing (ICSLP)*, vol. 2, pp. 883–886.

Lee, K.-F., Hon, H.-W., and Reddy, R. (1990). An overview of the SPHINX speech recognition system. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *38*(1):35–45.

Lee, T., Lo, W.K., Ching, P.C., and Meng, Helen. (2002a). Spoken language resources for Cantonese speech processing. *Speech Communication*, *36*:327–342.

Lee, T., Lau, W., Wong, Y.W., and Ching, P.C. (2002b). Using tone information in Cantonese continuous speech recognition. *ACM Transactions on Asia Language Information Processing*, *1*(1):83–102.

Levinson, S.E. (1986). Continuously variable duration Hidden Markov Models for automatic speech recognition. *Computer Speech and Language*, *1*:29–45.

Lin, T. and Wang, L.J. (1992). *Yu Yin Xue Jiao Cheng (in Pinyin)*. Beijing University Publishing.

Linguistic Society of Hong Kong (LSHK). (2002). *Hong Kong Jyut Ping Character Table*, 2nd ed. Linguistic Society of Hong Kong.

Peng, G. (2002). Reliability index guided prosody modeling in speech recognition. Ph.D. Dissertation, City University of Hong Kong.

Potisuk, S., Harper, M.P., and Gandour, J. (1999). Classification of Thai tone sequences in syllable-segmentated speech using the analysis-by-synthesis method. *IEEE Transactions on Speech and Audio Processing*, *7*(1):95–102.

Rabiner, L.R. (1984a). On the application of energy contours to the recognition of connected word sequence. *AT&T Bell Laboratories Techinical Journal*, *63*(9):1981–1995.

Rabiner, L.R. (1984b). On the performance of isolated word speech recognizers using vector quantization and temporal energy contours. *AT&T Bell Laboratories Techinical Journal*, *63*(7):1245–1260.

Rabiner, L.R. (1989). High performance connected digit recognition using Hidden Markov Models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *37*(8):1214–1225.

Ramesh, P. and Wilpon, J.G. (1992). Modeling state durations in Hidden Markov Models for automatic speech recognition. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 381–384.

Russell, M.J. and Moore, R.K. (1985). Explicit modeling of state occupancy in Hidden Markov Models for automatic speech recognition. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5–8.

Shen, X.-N. (1990). Tonal coarticulation in Mandarin. *Journal of Phonetics*, *18*:281–295.

Talkin, D. (1995). A robust algorithm for pitch tracking. In W.B. Kleijn and K.K. Paliwal (Eds.), *Speech Coding and Synthesis*. Amsterdam and New York: Elsevier, chapter 14, pp. 495–518.

Wang, W.S.-Y. (1973). The Chineese language. *Scientific American*, *228*:50–63.

Wang, W.S.-Y. and Li, K.-P. (1967). Tone 3 in Pekinese. *Journal of Speech and Hearing Research*, *10*(3):629–636.

Wilpon, J.G., Lee, C.-H., and Rabiner, L.R. (1991). Improvements in connected digit recognition using higher order spectral and energy features. *Proceedings of the International Conference on Acoustics, Speech, and Signal Procesing (ICASSP)*, vol. 1, pp. 349–352.

WiseNews. (2001). [Online]. Available: http://libwisenews.wisers.net.

Wu, Z.-J. (1984). Tone sandhi of tri-syllabic words in Mandarin. *Journal of Chinese Linguistics*, *2*:70–92.

Xu, Y. (1994). Production and perception of coarticulated tones. *Journal of the Acoustical Society of America (JASA)*, *95*(4):2240–2253.

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics*, *25*:61–83.

Zhang, B., Liu, J., Peng, G., and Wang, W.S.-Y. (1999). A high performance Mandarin digit recognizer. *Proceedings of the Fifth International Symposium on Signal Processing and its Applications (ISSPA)*, vol. 2, pp. 629–632.