Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 45 (2005) 49–62
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Tone recognition of continuous Cantonese speech based on support vector machines

Gang Peng [a,*], William S.-Y. Wang [b]

[a] Department of Electronic Engineering, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong
[b] Department of Electronic Engineering, Chinese University of Hong Kong, Shatin, New Territories, Hong Kong

Received 11 July 2003; received in revised form 23 September 2004; accepted 23 September 2004

Abstract

Tone is an essential component for word formation in all tone languages. It plays a very important role in the transmission of information in speech communication. In this paper, we look at using support vector machines (SVMs) for automatic tone recognition in continuously spoken Cantonese, which is well known for its complex tone system. An adaptive log-scale 5-level F_0 normalization method is proposed to reduce the tone-irrelevant variation of F_0 values. Furthermore, an extended version of the above normalization method that considers intonation is also presented. A tone recognition accuracy of 71.50% has been obtained in a speaker-independent task. This result compares favorably with the results reported earlier for the same task. Considerable improvement has been achieved by adopting this tone recognition scheme in a speaker-independent Cantonese large vocabulary continuous speech recognition (LVCSR) task.

Keywords: Tone language; F_0 normalization; Support vector machines; Tone recognition; Automatic speech recognition

1. Introduction

Tone is an essential component of tone languages, and is used to build words much as consonants and vowels do. For instance, in Cantonese, the syllable /si/, when pronounced with a high level pitch pattern, means "poetry"; with a high rising pattern, the meaning is "history"; with a middle level pattern, the meaning is "to try"; with a low falling pattern, the meaning is "time"; with a low rising pattern, the meaning is "city"; with a low level pattern, the meaning is "yes" (Wang and Cheng, 1987). So speech recognition of tone languages depends not only on the articulatory composition but also on the tone patterns.

During the last two decades, many approaches have been proposed for tone recognition. Hidden

* Corresponding author. Tel.: +852 2194 2632; fax: +852 2788 7791.
E-mail addresses: gpeng@ee.cityu.edu.hk (G. Peng), wsywang@ee.cuhk.edu.hk (W.S.-Y. Wang).

0167-6393/$ - see front matter
doi:10.1016/j.specom.2004.09.004

Markov models (HMMs) (Yang, 1988; Chen et al., 1987; Lee, 1997; Zhang, 2000; Wang, 2001), neural networks (Lee et al., 1995; Chen and Wang, 1995; Emonts and Lonsdale, 2003) and Fujisaki's model (Wang et al., 1990; Potisuk et al., 1999) have been applied to recognize tones in tone languages, such as Mandarin, Cantonese and Thai. For isolated tone recognition, very high recognition accuracy has been obtained (Yang, 1988; Emonts and Lonsdale, 2003). However, for tone recognition in continuous speech, although relatively high tone recognition accuracy has been achieved in Chen and Wang (1995) and Zhang (2000) and Potisuk et al. (1999) for Mandarin and Thai, respectively, manual segmentation was done before training the tone models, which is not suitable for automatic speech recognition (ASR). For automatic segmentation, a recognition score of 72.92% has been reported in Qian et al. (2003) for Cantonese with phonological constraints. However, without phonological constraints, lower recognition scores of 70.1% (Cao et al., 2000) and 66.4% (Lee et al., 2002a) have been reported for Mandarin and Cantonese, respectively.

In this paper, an adaptive log-scale 5-level F_0 normalization scheme will be presented to reduce tone-irrelevant variations. The intonation effect on tones will be further incorporated into the tone normalization scheme. Based on these schemes, considerable improvement of Cantonese tone recognition has been obtained. Furthermore, possible reasons for the severe confusion between some tone pairs will be discussed. When incorporating this tone recognition system into Cantonese LVCSR via the Parallel Tone Score Association (PTSA) (Peng and Wang, 2004) method, experimental results show that the relative character error rate was reduced by 5.1% compared with the recognition results with the baseline tone recognition scheme.

In the next section, the phonology and phonetics of Cantonese will be briefly introduced. The automatic extraction of short-term based tone features, the baseline F_0 normalization scheme, and the adaptive F_0 normalization schemes will be described in Section 3. The SVM based tone recognition method will be given in Section 4. Then the experimental results .will be presented in Section 5. Finally, conclusions will be drawn in Section 6.

2. Cantonese phonology and phonetics

Each Cantonese utterance can be viewed as a concatenation of monosyllables, each of which corresponds to a Chinese character. In Cantonese, there are a total of 1,761 tonal syllables; without considering tonal differences, there are about 625 base syllables (Linguistic Society of Hong Kong, 2002).

2.1. Syllabic structure

As shown in Fig. 1, each syllable of Cantonese can be divided into an optional Initial and an obligatory Final. The Initial is a consonant, while the Final consists of a nucleus complex with or without an ending, which may be a nasal (-m, -n, -ŋ) or a stop (-p, -t and -k).

Table 1 lists the Initials of Cantonese. The 19 Cantonese Initials are labeled with JyutPing (Linguistic Society of Hong Kong, 2002) spelling and the International Phonetic Alphabet (IPA) enclosed between slant lines. Table 2 lists the Finals of Cantonese. There are 53 Finals in Cantonese.

2.2. Lexical tones

Cantonese has a rich inventory of tones. Traditionally, it is said there are nine tones in Cantonese, as idealized in text-book fashion in Fig. 2. However, Fig. 3 displays time-aligned F_0 contours of the nine tones produced by a male subject who speaks native Hong Kong Cantonese. Tones 7, 8 and 9 are short tones because they alone have stop endings; these are called checked syllables. Since their F_0 values correspond to the long tones 1, 3

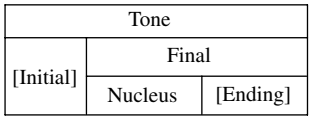

Fig. 1. Syllable structure of Cantonese. [ ] means that the enclosed element is optional (Wang, 1973).

Table 1
19 Initials of Cantonese in JyutPing and IPA

Cantonese			
JyutPing	IPA	Manner	Place
b	/p/	Plosive, unaspirated	Labial
d	/t/	Plosive, unaspirated	Alveolar
g	/k/	Plosive, unaspirated	Velar
gw	/k^w/	Plosive, unaspirated	Labial–velar
p	/p‘/	Plosive, aspirated	Labial
t	/t‘/	Plosive, aspirated	Alveolar
k	/k‘/	Plosive, aspirated	Velar
kw	/$k^{w‘}$/	Plosive, aspirated	Labial–velar
z	/ts/, /tʃ/	Affricate, unaspirated	Alveolar
c	/ts‘/, /tʃ‘/	Affricate, aspirated	Alveolar
s	/s/, /ʃ/	Fricative	Alveolar
f	/f/	Fricative	Labial–dental
h	/h/	Fricative	Glottal
l	/l/	Liquid	Lateral
m	/m/	Nasal	Labial
n	/n/	Nasal	Alveolar
ng	/ŋ/	Nasal	Velar
j	/j/	Glide	Alveolar
w	/w/	Glide	Labial

and 6, respectively, they are labeled according to their long tone counterparts. This is done in many transcription schemes, including that of the Linguistic Society of Hong Kong (LSHK), where only six distinct tones are labeled. There are totally five levels in Fig. 2. And the idea of using five levels to transcribe tones was first proposed in Chao (1930), which has been generally accepted for describing tones in Chinese, and generalized to all tone languages of the world (Wang, 1967).

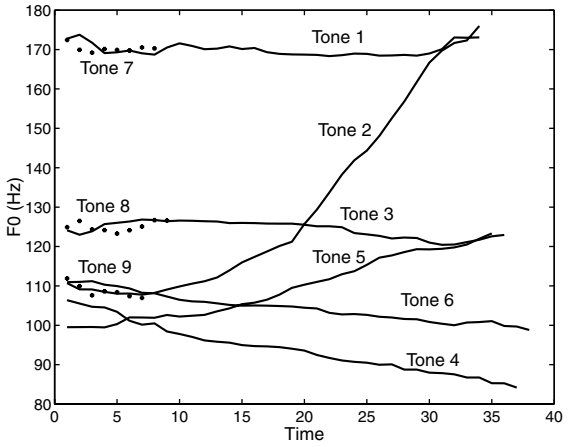

Fig. 3. F_0 contours of lexical tones of Cantonese uttered by a male speaker. The solid lines are for long tones on unchecked syllables, while the dotted lines are for short tones on checked syllables.

Table 2
53 Finals of Cantonese in JyutPing and IPA

	Cantonese
Vowel	i /i/, u /u/, yu /y/, e /ɛ/, oe /œ/, o /ɔ/, aa /a/
Vowel-nasal	im /im/, in /in/, ing /ɪŋ/, yun /yn/, un /un/, ung /ʊŋ/, eng /ɛŋ/, eon /∈n/, oeng /œŋ/, on /ɔn/, ong /ɔŋ/, am /ɐm/, an /ɐn/, ang /ɐŋ/, aam /am/, aan /an/, aang /aŋ/
Diphthong	ui /ui/, ei /ei/, eoi /∈y/, oi /ɔi/, ai /ɐi/, aai /ai/, iu /iu/, ou /ɔu/, au /ɐu/, aau /au/
Vowel-stop	ip /ip/, it /it/, ik /ɪk/, yut /yt/, ut /ut/, uk /ʊk/, ek /ɛk/, eot /∈t/, oek /œk/, ot /ɔt/, ok /ɔk/, ap /ɐp/, at /ɐt/, ak /ɐk/, aap /ap/, aat /at/, aak /ak/
Syllabic nasal	m /m̩/, ng /ŋ̩/

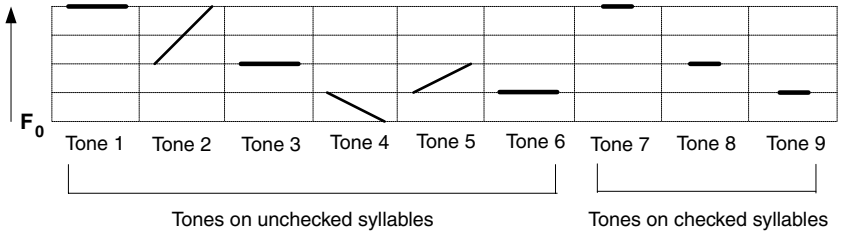

Fig. 2. Lexical tones of Cantonese as traditionally represented and labeled. Checked syllables end in the consonants /-p, -t, -k/; while other syllables are unchecked.

3. Feature extraction and normalization for tone recognition

Feature extraction is crucial for representing a speech signal in a compact and efficient manner for ASR. In this section, the tone feature extraction scheme and three normalization schemes are introduced.

3.1. Feature extraction for tone recognition

The tone of a syllable is mainly determined by its F_0 contour. The duration and energy are also related to the tone. Furthermore, tonal coarticulation effects, both carryover and anticipatory, have been studied extensively in Shen (1990) and Xu (1997) for Mandarin. For tone recognition in continuous speech, including contextual information from the neighboring tones improves the recognition accuracy (Cao et al., 2000; Lee et al., 2002a).

For a given syllable, a tone-related feature vector, called a token, consists of the following 20 features in our tone recognition scheme.

(1) Duration of the F_0 contour of the target syllable; F_0 values at both the 1/3 and 2/3 time points of each of the three uniformly divided linearly-fitted F_0 sub-contours; the means of the three corresponding log-energy sub-contours.
(2) The same three features (i.e., two F_0 values, mean of the log-energy) 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.
(3) Log-energy and duration of unvoiced/silent segments both before and after the F_0 contour of the target syllable.

As illustrated in Fig. 4, the 10 features in (1) are all extracted from the target syllable; the six features in (2) are used to consider the tonal coarticulation effect from the neighboring tones, while the four features in (3) are used to implicitly represent the degree of mutual influence between the target tone and its neighboring tones. This feature selection scheme is similar to Chen and Wang (1995). In our scheme, their *slope* feature of each linearly-fitted F_0 sub-contour is discarded, and their average F_0 (F_0 mean) is replaced by two F_0 values for each linearly-fitted F_0 sub-contour. We make these changes because in Mandarin, the *slope* feature is distinct for differentiate tones, but in Cantonese, the height of the F_0 contour becomes much more important because there are several level tones in Cantonese.

3.2. F_0 and energy normalization

Both the F_0 and energy are extracted frame by frame, with the same frame length and the same shift between successive frames as in the setup of feature extraction for training acoustic HMMs. F_0 is extracted with the autocorrelation method in Praat (Boersma and Weenink, 2001). The range of F_0 values varies dramatically among different speakers and also varies from time to time for each speaker. The dynamic range of signal energy may

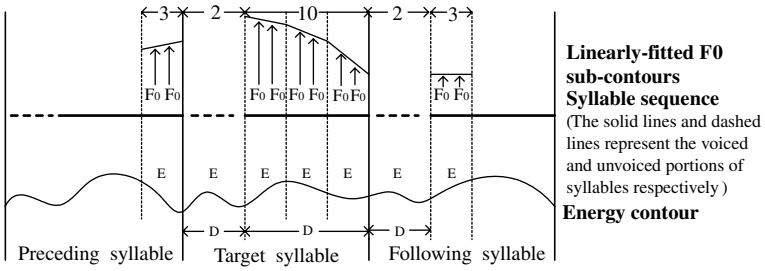

Fig. 4. Schematic diagram of the 20 tone features. *E* and *D* represent energy and duration, respectively. The numbers in the upper part of this figure indicate the numbers of features extracted from the corresponding speech segments.

vary significantly from utterance to utterance, and even within the same utterance. In Lee et al. (2002a), an effective method, called MWN (moving window normalization), has been proposed for Cantonese tone feature normalization, where the window extends to two preceding tones and four succeeding tones. Moreover, the average syllable duration of CUSENT database (Lee et al., 2002b), which is a read speech corpus and will be described Section 5, is about 0.25 s. Thus, we chose the normalization window extending to the past 0.5 s and the future 1 s of the target syllable. (The effect of different windows sizes on tone identification will be discussed in the discussion part of Section 5.) We make this change, using duration rather than the number of syllables to avoid the dependence on syllable boundaries during normalization. Furthermore, the log-scale transformation of parameters has been generally adopted in the framework of speech recognition (Rabiner and Juang, 1993). So a log-scale 5-level transformation will be used for the F_0 parameters according to:

$$F_0'(i) = \frac{\log_{10}(F_0(i)/\mathrm{Min})}{\log_{10}(\mathrm{Max}/\mathrm{Min})} \times 4 + 1. \quad (1)$$

Herein, three schemes are involved in determining the Max and Min. This transformation will be examined in the discussion part of Section 5.

1. Basic normalization scheme: Min and Max represent the minimum and maximum F_0 values within the above normalization window, respectively.
2. Adaptive normalization scheme: a reservoir which holds up to 4000 F_0 values is used to track the dynamic F_0 range of a speaker. A description of the scheme follows:
   (a) The reservoir is first initialized to empty when processing an utterance from a new speaker (we need not know any other information about the new speaker).
   (b) The F_0 values are inserted into the reservoir in ascending order. If the number of F_0 values exceeds the capacity (herein 4000) of the reservoir, half of the F_0 values in the reservoir are deleted alternately.
   (c) The lowest 5% and the highest 5% of the stored F_0 values are deleted in order to ignore outliers when determining the F_0 range of the speaker. This limits the impact of noise (e.g., sub and integral multiples of the true F_0 values) in the F_0 values. The corresponding tone recognition accuracy is reduced by about 10% when the outlier F_0 values are *not* excluded. Then the low and high ends of the remaining F_0 values are used for the Min and Max in Eq. (1), respectively, of the F_0 range of the speaker.
3. Extended adaptive normalization scheme: Multiple reservoirs which each hold up to 2000 F_0 values are used to track the dynamic F_0 range of a speaker according to different time courses. The first reservoir is used to track the F_0 range of the first second of the utterances; the second reservoir will be used to track the F_0 range of the second sec. of the utterances; and so on. The number of the reservoirs depends on the duration of the longest utterance. A description of the scheme follows:
   (a) The reservoirs are first initialized as empty when processing an utterance from a new speaker (we need not know any other information about the new speaker).
   (b) The F_0 values are inserted into the reservoirs in ascending order according to the following rules. The first reservoir is set as the default reservoir. When there are not enough F_0 values in this reservoir, all new coming F_0 values are inserted into this reservoir until the number of the stored F_0 values in this reservoir reaches the *Usable Point*(UP) (herein 400). After that, the F_0 values of the syllables whose starting time falls in the first second will be inserted into the first reservoir; the F_0 values of the syllables whose starting time falls in the second sec. will be inserted into the second reservoir; and so on. If the number of the F_0 values exceeds the capacity (herein 2000) of any reservoir, half of the F_0 values in this reservoir are deleted alternately.
   (c) During normalization, when the starting time of the target syllable falls in the nth s, we first check the number of the stored F_0 values in the nth reservoir. If the number is no less than the UP, then this reservoir

is used to calculate the corresponding F_0 range. Otherwise, the previous reservoir is checked until the number of the stored F_0 values in this reservoir is no less than the UP or the first reservoir is reached. The lowest 5% and the highest 5% of the stored F_0 values in the found reservoir are deleted in order to ignore outliers when determining the F_0 range of the speaker. Then the low and high values of those remaining correspond to the Min and Max in Eq. (1), respectively, of the F_0 range of the speaker during the corresponding time course.

Please note: the schemes 1, 2 and 3 of the tone normalization correspond to the schemes 1, 2 and 3 of the tone recognition systems, respectively.

The short-time log-energy is formulated as

$$E = 10\log_{10}[R(0)], \tag{2}$$

where $R(0)$ is the zeroth-order autocorrelation coefficient of the discrete time signal of speech. Then the log-scale energy is further re-scaled by the average log-energy within the normalization window.

3.3. Discussion

As is well known, there are large differences in the F_0 range among different speakers. The second normalization scheme is proposed to capture the dynamic F_0 range of the speakers. F_0 declination over an utterance is a common phenomenon in many languages including Cantonese, Mandarin and Thai (Ohala, 1978; Li et al., 2002; Kochanski and Shih, 2003; Potisuk et al., 1999). Fig. 5 depicts the effect of the declination intonation pattern on several tones. The second Tone 6 is lower than the first Tone 6; the second Tone 5 is lower than the first Tone 5; and the second Tone 3 is lower than the first Tone 3. The third normalization scheme is proposed to account for the declination effect on tone recognition.

Tone feature normalization is crucial for tone recognition. MWN (Lee et al., 2002a) has been proposed for Cantonese continuous tone normalization. A similar method has been adopted in our basic normalization scheme. The second and third normalization schemes have been proposed for better tone recognition performance.

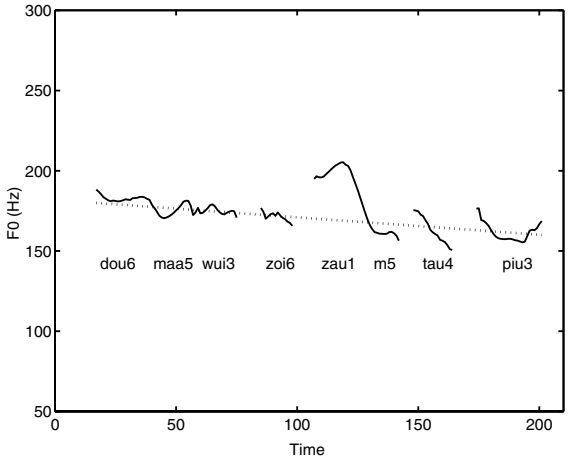

Fig. 5. F_0 contour of a Cantonese utterance by a female speaker. The numbers in the syllables refer to Cantonese tones as defined by the official spelling system, JyutPing.

4. Tone recognition based on SVMs

Alongside with normalization, tone recognition is also a problem of classification, that is, of assigning unknown tone tokens to a finite set of classes or categories (tone labels). In this section, SVM-based binary-class classifiers will be first introduced. Then the method of constructing tone recognition system via binary-class classifiers will be presented.

Classifiers are typically optimized based on some form of error minimization. Given a set of ℓ labeled examples:

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell), \quad \mathbf{x}_i \in \Re^N, y_i \in \{1, -1\},$$

where $\mathbf{x}_\ell$ is the example and y_ℓ is the class label ($\mathbf{x}_\ell$ and y_ℓ correspond to tone tokens and tone labels in tone recognition problem, respectively).

And a class of functions:

$$S = \{f_\alpha\}, f_\alpha : \Re^N \to \{1, -1\},$$

where α is an index and S is the function set.

The optimal function, $f_{\alpha'}$ in S is to be found by minimizing some error functions. Typically, empirical error minimization is one of the most straightforward approaches when the goal is to

find the function, i.e., the parameter set, which minimizes the following error function (empirical risk):

$$R_{\mathrm{emp}}(\alpha) = \frac{1}{2\ell}\sum_{i=1}^{\ell}|1 - y_i f_\alpha(\mathbf{x}_i)|, \tag{3}$$

where α is an index for the parameter set, y_i is the class label and also the expected output while $\mathbf{x}_i$ is the given input, and ℓ is the size of the training data.

Minimizing the empirical error can minimize the classification error on the training data but usually does not generalize well. As shown in Fig. 6, the training examples are divided into two classes, in which the empty circles belong to class I and the black solid circles belong to class II. All the hyperplanes, C_0, C_1 and C_2, have zero empirical error, and thus achieve perfect classification. Suppose that C_2 is selected as the classifier and a new unknown sample, as shown in Fig. 6, identified by the rectangular box comes for classification. Obviously, the new sample should be reasonably classified as class I, but it will be wrongly recognized as class II by C_2. Among the three hyperplanes, C_0 is the optimal hyperplane because it maximizes the distance between the H_1 and H_2, thereby offering better generalization.

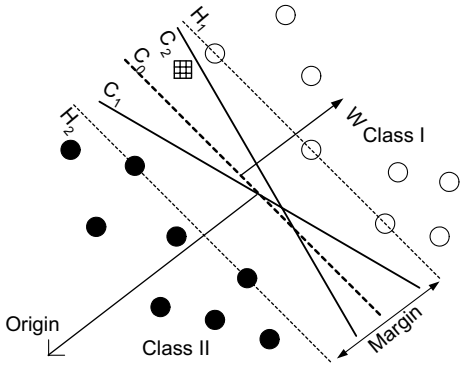

Fig. 6. Binary classification problem.

SVMs are based on structural risk minimization (SRM) where the aim is to learn a classifier that minimizes the bound of the expected error (Burges, 1998; Vapnik, 1995). Therefore, SVMs can guarantee the optimal solution of the target problem. As for the above example, the C_0 will be selected as the classification hyperplane by SVMs.

The power of SVMs lies in their ability to transform data to high dimensional space where the data can be separated using a linear hyperplane. SVMs have been successfully applied to many pattern recognition problems. Consequently, techniques of SVMs have been used to construct our tone classifiers.

SvmFu (Rifkin, 2001) was used to train a set of binary-class classifiers. By using the approach of error correcting output codes (ECOC) (Hastie and Tibshirani, 1998), the binary-class SVM classifiers are extended to perform multi-class classification. For instance, if there are six tones in Cantonese, then 15 (C_6^2) binary-class classifiers have to be trained; an ECOC matrix with six rows is given in Table 3.

For a given token $\mathbf{x}$, the 15 binary-class classifiers are applied to produce 15 hypotheses, h_1, $h_2,\ldots,h_{15}$. Then the class label of token $\mathbf{x}$ can be predicted by choosing the rth row of the ECOC matrix which is closest to $(h_1(\mathbf{x}), h_2(\mathbf{x}),\ldots,h_{15}(\mathbf{x}))$. Furthermore, a classification score can be assigned to each class label by

Table 3
ECOC matrix for a six-class problem, where 1/2 represents Tone 1 vs. Tone 2, 1/3 represents Tone 1 vs. Tone 3, etc.

One-versus-one (pairwise)														
1/2	1/3	1/4	1/5	1/6	2/3	2/4	2/5	2/6	3/4	3/5	3/6	4/5	4/6	5/6
1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
−1	0	0	0	0	1	1	1	1	0	0	0	0	0	0
0	−1	0	0	0	−1	0	0	0	1	1	1	0	0	0
0	0	−1	0	0	0	−1	0	0	−1	0	0	1	1	0
0	0	0	−1	0	0	0	−1	0	0	−1	0	−1	0	1
0	0	0	0	−1	0	0	0	−1	0	0	−1	0	−1	−1

$$D(M(r), \mathbf{h}(\mathbf{x})) = \sum_{s=1}^{l} e^{M(r,s)h_s(\mathbf{x})}, \quad (4)$$

where M is the code matrix, $M(r)$ is the rth row of the code matrix, and l is the number of columns of the code matrix. The number of rows of the ECOC matrix is equal to the number of tone classes.

5. Experimental results

In this section, the database, the acoustic models and language models will be briefly introduced. Then the experimental results of Cantonese tone recognition will be presented. Possible reasons for the severe confusions between some tone pairs will be discussed. Finally, contributions to speech recognition from different tone recognition schemes will be compared.

5.1. Database

The Cantonese database we used is CUSENT database (Lee et al., 2002b). In this database, 5100 training and another exclusive 600 test sentences were selected from five local newspapers of Hong Kong. The training sentences were evenly divided into 17 groups, each containing 300 unique sentences. Each group of sentences was read by four speakers (2F, 2M). Thus, a total of 20,400 ($300 \times 4 \times 17$) training utterances were obtained from 68 speakers. The 600 test sentences were divided into six groups. Each group was read by one male and one female speaker (not drawn from the population of the training speakers). The total number of test utterances is 1200. Table 4 summarizes information about the database. Please note corrupted recording utterances were excluded by the developers (Lee et al., 2002b), so the numbers of utterances for training and test were reduced from 20,400 and 1200 to 20,378 and 1198, respectively.

Table 4
Details of the CUSENT database

Properties	Training data	Test data
Number of speakers	68(34F,34M)	12(6F,6M)
Number of syllables	215,604	11,677
Number of sentences	20,378	1198

5.2. Acoustic modeling and language modeling

The acoustic models consist of context-dependent Initial–Final models (tri-phone models), in which each Initial model has three emitting states, while a Final model has either three or five emitting states, depending on its articulatory composition. The acoustic feature vector has a total of 39 components, including 12 MFCCs, energy, their first-order derivatives and second-order derivatives. The HMMs were trained with the training set of the CUSENT database. A decision tree-based clustering method was used to facilitate sharing of model parameters.

After clustering, HMM states of some triphone models were tied and the number of HMM parameters reduced dramatically. The Gaussian mixture number of the HMMs was then increased gradually by mixture splitting to model the speaker variation effect. As shown in Table 5, when the number of mixtures increases from eight to nine, the performance improvement is not obvious, so the eight-mixture system has been selected for further processing. Please note that this result (79.09%) outperforms the biphone system (73.1%) using the same database (Lee et al., 2002b).

The language models, character-based bigrams and trigrams, have been built with 3927 character entries, which cover 99.99% of the training Cantonese text corpus. The training text contains about 150 million Chinese characters from Wise-

Table 5
Performance of the triphone models with different number of mixtures, where 'base syllable accuracy' is the syllable accuracy without considering tonal differences

Number of mixtures	Base syllable accuracy (%)
1	73.12
2	75.62
3	76.76
4	77.34
5	78.12
6	78.51
7	78.89
8	79.08
9	79.18

News (2001). Using the test data of CUSENT database without tone information, character accuracies of 72.40% and 80.90% were obtained for bigrams and trigrams, respectively. This eight-mixture system together with the character-based trigrams serves as the baseline system in this paper.

5.3. Experimental results of tone recognition

To obtain the training and test tone tokens for tone recognition, forced alignment by the baseline HMMs was applied to obtain Initial–Final segmentation for all training and test utterances.

Tone tokens extracted from 5992 training utterances of the training set of CUSENT, from 20 (10 M, 10 F) randomly selected speakers, were used to train the tone classifiers. Then tokens extracted from all utterances in the test set of CUSENT were used to evaluate the performance of the tone classifiers.

The confusion matrices from the three tone normalization schemes are shown in Tables 6–8, respectively. Recall from Section 2, the number

Table 6
Confusion matrix of tone recognition with tone normalization scheme 1

	Recognized tone						Total tokens	Accuracy (%)
	Tone 1	Tone 2	Tone 3	Tone 4	Tone 5	Tone 6		
Tone 1	2201	38	141	6	7	38	2431	90.54
Tone 2	79	1151	65	44	82	73	1494	77.04
Tone 3	331	67	927	87	26	480	1918	48.33
Tone 4	33	49	124	1334	26	517	2083	64.04
Tone 5	17	221	72	81	250	174	815	30.67
Tone 6	144	85	524	305	44	1298	2400	54.10
Overall								64.28

Table 7
Confusion matrix of tone recognition with tone normalization scheme 2

	Recognized tone						Total tokens	Accuracy (%)
	Tone 1	Tone 2	Tone 3	Tone 4	Tone 5	Tone 6		
Tone 1	2259	30	101	6	3	32	2431	92.92
Tone 2	83	1176	53	50	61	71	1494	78.71
Tone 3	215	74	1128	57	29	415	1918	58.81
Tone 4	21	37	60	1575	20	370	2083	75.61
Tone 5	7	213	55	82	259	199	815	31.87
Tone 6	83	91	531	239	48	1408	2400	58.67
Overall								70.06

Table 8
Confusion matrix of tone recognition with tone normalization scheme 3

	Recognized tone						Total tokens	Accuracy (%)
	Tone 1	Tone 2	Tone 3	Tone 4	Tone 5	Tone 6		
Tone 1	2263	32	105	9	2	20	2431	93.09
Tone 2	65	1190	50	43	64	82	1494	79.65
Tone 3	160	71	1156	43	31	457	1918	60.27
Tone 4	17	40	39	1589	15	383	2083	76.28
Tone 5	6	197	55	87	281	189	815	34.48
Tone 6	52	90	505	211	55	1487	2400	61.96
Overall								71.50

of actual tonal syllables, 1761, is less than half of the number of possible syllables, i.e., $6 \times 625 = 3750$. A simple phonological constraint can be applied to eliminate those recognition results that are not actual tonal syllables. Using this simple phonological constraint, our recognition score for the first normalization scheme increases from 64.28% to 69.39%; for the second normalization scheme our score increases from 70.06% to 75.31%; and for the third normalization scheme our score increases from 71.50% to 77.27%. Using the same database, i.e., CUSENT, a tone recognition accuracy of 66.4% has been reported in Lee et al. (2002a) based on HMMs; with the phonological constraint, an accuracy of 72.92% has been achieved by Qian et al. (2003) using overlapped di-tone gaussian mixture models (ODGMM) for the same database. Compared with these previously reported results, our algorithm using the third normalization scheme achieves a substantially greater accuracy.

5.4. Discussion

5.4.1. Discussion about the tone feature selection, transformation and window size selection

As stated in Section 1, many approaches have been proposed for tone recognition; an approach that used neural networks to recognize continuous Mandarin tone (Chen and Wang, 1995) is among the most successful. Furthermore, Chen and Wang's features can be adapted to SVMs. Hence, we have adopted their features except for changing their *slope* and F_0 means to two F_0 values for each linearly-fitted F_0 sub-contour, as discussed in Section 3. First, we compare the tone recognition accuracies of Chen and Wang's features with our different normalization methods. The advantage of moving windows normalization (MWN) can be observed in the first three rows of Table 9. Next, we compare our features with Chen and Wang's features for Cantonese tone recognition. As indicated in rows (3) and (4) in Table 9, our features outperform Chen and Wang's features by a considerable margin. It is likely that 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.

Table 9
Tone feature selection and transformation

Feature	Accuracy (%)
(1) Chen and Wang's features (without normalization)	57.80
(2) (1) + Utterance wide normalization	58.55
(3) (1) + MWN	60.01
(4) Our features + MWN	64.06
(5) (4) + Simple log-transformation	57.17
(6) (4) + 5-Level transformation (in linear-scale)	63.52
(7) (4) + 5-Level transformation (in log-scale)	64.28
(8) (4) + 3-Level transformation (in log-scale)	62.37
(9) (4) + 7-Level transformation (in log-scale)	64.18
(10) (4) + 9-Level transformation (in log-scale)	63.94

Several transformations of F_0 values have been examined. In row (4), the absolute F_0 values are transformed to relative ratios over the average F_0 value (as reference F_0 value) within the normalization window. But in row (5), the F_0 values are first transformed into log-scale. Comparing row (5) with row (4), this kind of simple log-transformation lowers the recognition performance. The possible reason is that this kind of log-transformation dramatically diminishes the differences between F_0 values. For instance, in linear-scale, 140/150 minus 120/150 is 0.1333, while in log-scale, $\log_{10}(140)/\log_{10}(150)$ minus $\log_{10}(120)/\log_{10}(150)$ is only 0.0308 (150 serves as the reference F_0 value). However, when using 5-level transformation, the transformation in log-scale does better than its counterpart, as are shown in rows (6) and (7) of Table 9. As shown in the last four rows of Table 9, 5-level transformation is better than other choices. Interestingly, this is consistent with the selection of number of levels for linguists to transcribe tones (Chao, 1930).

The window sizes of MWN in Table 9 always used 0.5 s left context and 1 s right context. In order to investigate the effect of different window sizes on tone identification, different window sizes have also been examined.

We first fix the length of right context to 1 s. From the left-most columns of Table 10, we observe that the accuracy decreases as the length of left context increases beyond 0.25 s, so we conserv-

Table 10
Tone recognition with different windows sizes

1 s of right context		0.5 s of left context	
Length of left context (s)	Tone recognition accuracy (%)	Length of right context (s)	Tone recognition accuracy (%)
0.25	64.30	0.50	63.80
0.50	64.28	0.75	64.29
0.75	64.04	1	64.28
1	63.99	1.25	64.23

atively chose 0.5 s as the length of left context. Then we fix the length of left context to 0.5 s. From the right-most columns of Table 10, we observe the accuracy decreases as the length of right context drops below 0.75 s or increases beyond 1 s. So we conservatively choose 1 s as the length of right context. This window size selection is consistent with that reported in Lee et al. (2002b), where number of syllables was used (recall the average syllable duration is 0.25 s).

5.4.2. Analysis of Cantonese tone recognition

With the help of the adaptive normalization scheme, the performance of the tone recognition system improves from 64.28% to 70.06%, i.e., a 16.18% relative error rate reduction. When the intonation effect is simply considered by using multiple reservoirs, a further 4.81% relative error rate reduction has been obtained. Please note that most utterances in CUSENT have a downdrift intonation pattern because of the reporting style characteristic of this database. However, the intonation patterns of spontaneous speech will be much more complex. Therefore, further research based on more complex intonation patterns will be needed for improving tone recognition accuracy of spontaneous speech.

As shown in the confusion matrices, the highest recognition accuracy has been obtained for Tone 1 (high level tone). Because it occupies the highest position of the Cantonese tone space; it is relatively easy to distinguish this tone from all of the other tones. Tones 3–6, which are at the middle and low regions of the tone space, are much more difficult to recognize correctly. Moreover, the numbers of Tone 2 and Tone 5 tokens, are much less (2309) than the numbers of tokens from other tones. This is not due to any bias in the database we used, but due to the normal distributional characteristics (Wang and Cheng, 1987). In these tone recognition experiments, the accuracy of Tone 5 is extremely low, which is not consistent with the results from previous perceptual studies of Cantonese tones (Fok, 1974; Man, 1992).

In order to investigate possible reasons why some tone pairs are so severely confused, an informal tone perception experiment has been carried out. Six native Hong Kong residents served as subjects. Syllables corresponding to the two tones most frequently confused by our best tone recognition system, i.e., Tone 3 and Tone 6, were extracted for two female and two male speakers. Two subsets of these syllables were used in this experiment: syllables with Tone 3 but mis-recognized as Tone 6, and syllables with Tone 6 but mis-recognized as Tone 3 by our best tone recognition system. The number of such syllables used for this perception experiment is 384. In total, there are 2304 (384×6) trials; in each trial, subjects are required to select the tone (Tone 3 or Tone 6) that they perceive, or, if they cannot identify the tone, to indicate that fact.

During the perception experiment, two selected utterances of each speaker, including all Cantonese tones, are played twice to subjects for reference. The syllable sets from each speaker are presented to the subjects in random order; each syllable is repeated twice with a separation of 1 s; the interval between successive syllables is 3 s. This perception experiment is carried out speaker by speaker. In our experiment, as shown in Table 11, there were 551 trials in which the subject made no decision; there were 948 correct decisions and 805 incorrect decisions; responses were identified as correct or not based on the transcription provided by the database.

Table 11
Results from perception experiment on Tone 3 and Tone 6 without preceding and following tonal context

	Tone 3	Tone 6	Total	Ratio (%)
Correct	504	444	948	41.15
Incorrect	312	493	805	34.94
Cannot decide	273	278	551	23.91

Table 12
Results from perception experiment on Tone 3 and Tone 6 with preceding and following tonal context

	Tone 3	Tone 6	Total	Ratio (%)
Correct	493	552	1045	68.03
Incorrect	185	217	402	26.17
Cannot decide	54	35	89	5.80

An additional perception experiment was carried out to examine the effect of tonal context on tone perception. In this experiment, we also preserved both the preceding and following tonal context. The number of trials was changed to 1536 (384×4). All other conditions are kept unchanged. The percentage of subjective uncertainty dropped dramatically from 23.91% to 5.8%. However, as shown in Table 12, the confusion between Tone 3 and Tone 6 was severe even with the help of the limited tonal context.

One possible reason for the confusion may be that these two tones are in the process of merging as a sound change in progress. (A merger occurs when two phonemes become one phoneme. This has happened, for instance, for two English vowels, resulting in the homophony of words such as 'meet' and 'meat'. In our case, the two phonemes are Tone 3 and Tone 6, which cannot be correctly distinguished by either our tone recognizers or by human subjects cf. Table 8 and Table 12 above). Merger can take place at least in two ways, i.e., by speaker and by word (Wang, 1969). Such a merger might also occur between other tone pairs. This is an interesting question which merits further study. Another possible reason is that these two tones are both located at the lower part of the Cantonese tone space (recall Figs. 2 and 3), which is much more crowded than the upper part. Consequently, lower recognition accuracies have been obtained for those lower tones, especially for Tone 3, Tone 5 and Tone 6. Among these three tones, both Tone 3 and Tone 6 are level tones, which make them even more prone to error. And this might be one of the reasons for these two tones to merge. One other possible reason for the severe confusion in this pair of tones might be segmentation errors produced by the forced alignment, a discrepancy between the segmentation of computers vs. humans, however, these errors would afect the accuracy in recognizing all six tones.

5.5. Incorporation tone knowledge into Cantonese LVCSR

The Parallel Tone Score Association (PTSA) method is adopted here to incorporate tone recognition into speech recognition (Peng and Wang, 2004). The basic idea of PTSA is to add a tonal contribution in parallel to 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 the syllable lattice, then it will be expanded to all possible tonal syllables (since not all base syllables co-occur with every tone, only possible tonal syllables will be considered). How to distribute the total tonal contribution to each tonal syllable is the crucial aspect of the tonal syllable lattice generation. Because only voiced frames have meaningful F_0 values, and F_0 is by far the most important manifestation of tones, we define the total tonal contribution over the voiced frames as:

$$S_{\text{Tone}} = C \times L, \tag{5}$$

where C is a language related constant, and L is the number of voiced frames of the target syllable. If S_{Tone} is equally distributed to each tonal syllable, then no tonal syllables will be preferred; consequently, an equal amount of tonal score will be added to each path of the lattice. But in PTSA, each tone is assigned a recognition score by the tone classifier based on the tone-related feature vector (tone token), and then S_{Tone} will be distributed to each tonal syllable proportionally according to its tonal recognition score.

All three schemes of tone normalization are evaluated for determining tonal contributions to Cantonese speech recognition. Experimental

Table 13
Performance of the integrated systems

Speaker (M: male, F: female)	Base syllable accuracy (%)	Character accuracy (%)				
		Without tone (%)	With scheme 1 (%)	Improvement (%)	With scheme 2/3 (%)	Further improvement (scheme 2/3) (%)
01 (M)	75.81	83.12	86.84	3.72	87.55/88.00	0.71/1.16
02 (M)	83.66	85.50	89.00	3.50	89.50/89.60	0.50/0.60
03 (M)	73.52	79.97	81.51	1.54	82.30/82.40	0.79/0.89
04 (M)	78.00	79.98	85.50	5.52	86.31/86.52	0.81/1.02
05 (M)	82.33	76.88	83.96	7.08	83.83/84.38	−0.13/0.42
06 (M)	77.78	78.13	83.76	5.63	84.26/84.93	0.50/1.17
07 (F)	83.86	80.23	83.39	3.16	83.30/83.60	−0.09/0.21
08 (F)	84.00	84.59	87.07	2.48	88.78/88.93	1.71/1.86
09 (F)	76.89	79.53	83.32	3.79	83.90/84.09	0.58/0.77
10 (F)	79.57	81.49	84.56	3.07	84.45/84.56	−0.11/0.00
11 (F)	75.52	80.39	84.07	3.68	84.98/84.88	0.91/0.81
12 (F)	78.24	82.35	86.06	3.71	86.02/86.27	−0.04/0.21
Overall	79.08	80.90	84.90	4.00	85.42/85.66	0.52/0.76

results are shown in Table 13. Considerable improvement has been achieved by incorporating tone information into speech recognition using scheme 1. When using scheme 2, although overall performance is improved from 84.90% to 85.42%, four out of 12 speakers show lower scores compared with using scheme 1. When using scheme 3, there is consistent improvement for all speakers. Considerable improvements are obtained for more than half of the speakers. Furthermore, the overall performance of the LVCSR system is improved from 84.90% with scheme 1 to 85.66% with scheme 3, i.e., a 5.1% relative reduction in the character error rate. As for the value of the language related constant C, defined in Eq. (5), the values 30, 40 and 42 were selected empirically for scheme 1, scheme 2 and scheme 3, respectively. Larger C values were chosen for the stronger tone recognition systems due to their greater tone recognition accuracies.

6. Conclusions

Several tone normalization schemes have been investigated in the present paper. The adaptive normalization method was found to reduce tone-irrelevant variation of F_0 values. As a result, tone recognition accuracy was significantly improved from 64.28% to 70.06%, i.e., a 16.18% relative error rate reduction. When the intonation effect is taken into consideration in a very simple way, tone recognition accuracy was further improved from 70.06% to 71.50%, resulting in another 4.81% relative error rate reduction. This accuracy of 71.50% compares favorably with the 66.4% result reported earlier for the same task (Lee et al., 2002a). Speech recognition results also verify the expectation that the more accurate the tone recognizer is, the greater its contribution is to the speech recognition accuracy. This result should encourage further investigation to more fully exploit the potential of tone information for improving the tone recognition. The normalization schemes investigated in this paper can also be applied to other tone languages, such as Mandarin and Thai, to achieve similar benefits in speech recognition.

Acknowledgments

The work described in this paper was supported by grants from City University of Hong Kong (Project No. 7001327, 9010001). The authors would like to thank Dr. James W. Minett, Prof. H.T. Thomas Lee, Prof. C.C. Cheng and other colleagues in the Language Engineering Laboratory for their invaluable advice. We also thank Prof. P.C. Ching and Prof. T. Lee of Chinese University of Hong Kong for providing us the database CU-SENT, which they compiled. We thank the editor

and the reviewers for their constructive help in improving our paper.

References

Boersma, P., Weenink, D., 2001. Praat: doing phonetics by computer. (Online). Available from <http://www.fon.hum.uva.nl/praat/>.

Burges, C., 1998. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, Vol. 2. Kluwer Academic Publishers, Boston, No. 2, pp. 121–167.

Cao, Y., Deng, Y.G., Zhang, H., Huang, T.Y., Xu, B., 2000. Decision tree based Mandarin tone model and its application to speech recognition. In: Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 3, pp. 1759–1762.

Chao, Y.R., 1930. A system of tone letters. Le Maître Phonétique 45, 24–27.

Chen, X.-X., Cai, C.-N., Guo, P., Sun, Y., 1987. A Hidden Markov model applied to Chinese four-tone recognition. In: Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 797–800.

Chen, S.H., Wang, Y.R., 1995. Tone recognition of continuous Mandarin speech based on neural networks. IEEE Trans. Speech Audio Process. 3 (2), 146–150.

Emonts, M., Lonsdale, D., 2003. A memory-based approach to Cantonese tone recognition. In: Proc. 8th European Conference on Speech Communication and Technology (EUROSPEECH), pp. 2305–2308.

Fok, Y.Y.C., 1974. A Perceptual Study of Tones in Cantonese. Centre of Asia Studies, University of Hong Kong, Hong Kong.

Hastie, T., Tibshirani, R., 1998. Classification by pairwise coupling. In: Jordan, M.I., Kearns, M.J., Solla, S.A. (Eds.), Advances in Neural Information Processing Systems. MIT Press.

Kochanski, G., Shih, C.-L., 2003. Prosody modeling with soft templates. Speech Comm. 39 (3-4), 311–352.

Lee, L.S., 1997. Voice dictation of Mandarin Chinese. IEEE Signal Process. Mag. 14 (4), 63–101.

Lee, T., Ching, P.C., Chan, L.W., Cheng, Y.H., Mak, B., 1995. Tone recognition of isolated Cantonese syllables. IEEE Trans. Acoust. Speech Signal Process. 3 (3), 204–209.

Lee, T., Lau, W., Wong, Y.W., Ching, P.C., 2002a. Using tone information in Cantonese continuous speech recognition. ACM Trans. Asian Language Info. Process. 1 (1), 83–102.

Lee, T., Lo, W.K., Ching, P.C., Meng, H., 2002b. Spoken language resources for Cantonese speech processing. Speech Comm. 36 (3–4), 327–342.

Li, Y.-J., Lee, T., Qian, Y., 2002. Acoustical F0 analysis of continuous Cantonese speech. In: Proc. Internat. Symposium on Chinese Spoken Language Processing (ISCSLP), pp. 127–130.

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

Man, C.H.V., 1992. An acoustic study of the effects of sentential focus on Cantonese tones, Master's thesis, University of Victoria, British Columbia, Canada.

Ohala, J.J., 1978. Production of tones. In: Fromkin, V.A. (Ed.), Tone: A Linguistic Survey. Academic Press, New York, pp. 5–50.

Peng, G., Wang, W.S.-Y., 2004. Parallel tone score association method for tone language speech recognition. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP).

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

Qian, Y., Lee, T., Li, Y.-J., 2003. Overlapped di-tone modeling for tone recognition in continuous Cantonese speech. In: Proc. 8th European Conf. on Speech Communication and Technology (EUROSPEECH), pp. 1845–1848.

Rabiner, L.R., Juang, B.-H., 1993. Fundamentals of Speech Recognition. PTR Prentice Hall, Englewood Cliffs, NJ.

Rifkin, R., 2001. SvmFu documentation. (Online). Available from <http://five-percent-nation.mit.edu/SvmFu/index.html>.

Shen, X.N., 1990. Tonal coarticulation in Mandarin. J. Phonetics 18, 281–295.

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, New York, USA.

Wang, C., 2001. Prosodic Modeling for improved speech recognition and understanding. PhD dissertation, MIT.

Wang, C.F., Fujisaki, H., Chen, S.H., 1990. The four tones recognition of continuous Chinese speech. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP), Vol. 6, pp. 221–224.

Wang, W.S.-Y., 1967. Phonological features of tone. Int. J. Amer. Linguistics, 93–105.

Wang, W.S.-Y., 1969. Competing change as a cause of residue. Language 45, 695–708.

Wang, W.S.-Y., 1973. The Chinese language. Sci. Amer. 228, 50–63.

Wang, W.S.-Y., Cheng, C.C., 1987. Middle Chinese tones in modern dialects. In: Channon, R., Shockey, L. (Eds.), To Honor Ilse Lehiste. Foris Publishers.

WiseNews, 2001. (Online). Available from <http://libwisenews.wisers.net>.

Xu, Y., 1997. Contextual tonal variations in Mandarin. J. Phonetics 25, 61–83.

Yang, W.-J., 1988. Hidden Markov Model for Mandarin lexical tone recognition. IEEE Trans. Acoust. Speech Signal Process. 36, 988–992.

Zhang, J.-S., Hirose K., 2000. Anchoring hypothesis and its application to tone recognition of Chinese continuous speech. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pp. 1419–1422.