Correspondence

Tone Recognition of Isolated Cantonese Syllables

Tan Lee, P. C. Ching, L. W. Chan, Y. H. Cheng, and Brian Mak

Abstract—**Tone identification is essential for the recognition of Chinese language, specifically for Cantonese which is well known of being very rich in tones. This paper presents an efficient method for tone recognition of isolated Cantonese syllables. Suprasegmental feature parameters are extracted from the voiced portion of a monosyllabic utterance and a three-layer feedforward neural network is used to classify these feature vectors. Using a phonologically complete vocabulary of 234 distinct syllables, the recognition accuracy for single-speaker and multispeaker is given by 89.0% and 87.6% respectively.**

I. Introduction

The Chinese language is well known for being monosyllabic and tonal. The pronunciation of a Chinese character is completely determined by a single syllabic sound associated with a specific lexical tone. Cantonese, a commonly used dialect in Southern China, is particularly rich in tones. There are nine distinctive tones in contemporary Cantonese speech whereas there are only five tones in the official Chinese language, Mandarin (Putonghua). Speech recognition of tonal languages depends not only on the phonetic composition but also on the lexical tone pattern. Tones in monosyllabic languages are regarded as a feature of pitch movement within a syllable. Different tones can be identified and separated from each other by the static and temporal behavior of pitch. Several methods have been proposed for the recognition of Mandarin tones [1], [2] and most of them utilize the temporal variation of pitch as the main discrimination feature. These methods, however, are less successful for Cantonese because the tone systems of Mandarin and Cantonese are quite different in nature. The identification of a Cantonese tone relies on both the pitch contour and the relative pitch level. Furthermore, for separation of the so-called entering tones from the others, the duration also plays an important role as well [3].

In this paper an efficient method for tone recognition of isolated Cantonese syllables is proposed (Fig. 1). Five suprasegmental feature parameters are obtained for each Cantonese syllable, three of them are used to characterize the pitch contour of the syllable while the remaining two distinguish entering tones from nonentering tones. A normalization procedure is used to eliminate undesirable time variation and speaker variation of these feature parameters. A multilayer feedforward neural network is then applied to classify the normalized feature sets. Standard backpropagation algorithm is employed for supervised training of the network [4]. By activating one of its output neurons, the neural network recognizes the lexical tone of the input utterance.

Manuscript received July 19, 1993; revised November 10, 1994. This work was supported by a grant from the Croucher Foundation. The associate editor coordinating the review of this paper and approving it for publication was Dr. Amro El-Jaroudi.

T. Lee, P. C. Ching, and B. Mak are with the Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong.

L. W. Chan is with the Department of Computer Science, The Chinese University of Hong Kong, Hong Kong.

Y. H. Cheng is with Motorola Semiconductor Hong Kong Ltd., Hong Kong.

IEEE Log Number 9410226.

In the next section, a brief introduction to the phonology of Cantonese will be given and the acoustic features of the nine lexical tones will be identified. In Section III, the extraction of suprasegmental features for tone identification will be explained, which is followed by a description of the neural network classifier in Section IV. Recognition tests for both single-speaker and multispeaker have been performed and the results are shown in Section V.

II. Tone System of Cantonese Dialect

The Cantonese dialect studied in this paper is also referred to as the Guangzhou dialect. It is the most popular dialect in the group of multifarious "Yue" dialects. For some geographical and historical reasons, the term "Cantonese" has been employed to represent the standard or the norm of "Yue" dialects [3], [5]. A Cantonese syllable can be divided into an Initial and a Final (Fig. 2). The Initial is optional and if exists, can be either a glide, a nasal or an unvoiced consonant. The Final is made up of a syllabic segment and an optional ending. The syllabic segment is usually a vowel while the ending can be either a glide, a nasal or a glottalized stop consonant. There are totally 20 Initials, including the zero initial, and 53 Finals in Cantonese. If tone is ignored, there are about 580 phonemic combinations of these Initials and Finals in Cantonese.

The introduction of lexical tones expands the syllabary of Cantonese to about 1600 different syllable sounds [6], in which the distribution of different tone groups ranges from 68 to 280 (Table I). As depicted in Fig. 3, the nine tones of Cantonese are categorized into two groups, namely the nonentering group (tone 1–6) and the entering group (tone 7–9) [7]. The term "entering" designates a traditional Chinese tone category which refers to all syllables with stop endings. Cantonese is one of the few existing Chinese dialects that preserve entering tones.

The nonentering tone group can be further divided into upper series (tone 1, 3, 5) and lower series (tone 2, 4, 6), according to their relative pitch levels. Each of the upper and lower series is composed of a level tone, a rising tone and a going tone. Again, these are traditional Chinese phonological terminologies and they represent the tone categories of the ancient Chinese but not the tone realization of contemporary Cantonese. It is noted that only rising tones have temporally rising pitch contours, while all level tones and going tones have either flat or slightly falling pitch profiles.

The entering tones, on the other hand, do not have their own patterns of pitch movement. The most distinctive feature of entering tone syllables is that they always end with stop consonants, such as /*p*/, /*t*/ and /*k*/, and thus they are much shorter in duration. In normal speech, the duration of an entering tone is roughly 1/3 of that of a nonentering tone [8].

III. Suprasegmental Features for Tone Recognition

A. Pitch-Related Features

According to Chao [9], the exact temporal pitch variation is not absolutely essential for tone recognition of Chinese syllables from the phonological point of view. However, the six nonentering tones of Cantonese differ acoustically from each other either by the trend of their pitch contours or by their relative pitch levels. The tone

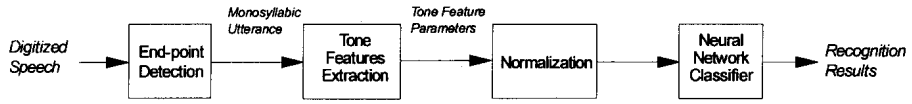

Fig. 1. A tone recognition system for isolated Cantonese syllables.

TABLE I
THE DISTRIBUTION OF LEXICAL TONES IN CANTONESE SYLLABARY

	Tone Number	Tone Name	No. of Syllables	Percentage
Non-Entering Tones	1	Upper Level	280	17.6 %
	2	Lower Level	217	13.6 %
	3	Upper Rising	233	14.7 %
	4	Lower Rising	143	9.0 %
	5	Upper Going	240	15.1 %
	6	Lower Going	212	13.3 %
Entering Tones	7	Upper Entering	68	4.3 %
	8	Middle Entering	101	6.4 %
	9	Lower Entering	97	6.1 %
		Total :	1591	100.0 %

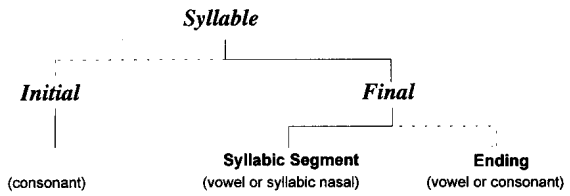

Fig. 2. The phonetic structure of a Cantonese syllable (dashed line means optional component).

3 (upper rising) and tone 4 (lower rising) syllables can be easily identified by the overall pitch rising while other nonentering tones and entering tones do not have such property. Furthermore, the lower tones usually have lower pitch values than the upper tones, and the difference in pitch level is also noticeable among the entering tones. Therefore, three parameters, namely an initial pitch, a final pitch and a pitch rising index, are derived from the pitch profile of each syllable to characterize the pitch level and the pitch movement for the identification of the nine Cantonese lexical tones.

To deal with different utterances with different duration, a time-aligned pitch profile is used [10]. We divide the voiced portion of a syllable evenly into 16 segments, and for each segment a pitch value is obtained by using the modified 3-level center clipping algorithm (Sondhi [11]). Thus the pitch profile of each and every syllable has the same dimension of 16. Given a pitch profile $\{P(1), P(2), \cdots, P(i), \cdots, P(16)\}$, the initial pitch and the final pitch, that represent the absolute pitch levels at the beginning and at the end of an utterance respectively, are defined as follows

$$P_I = \frac{P(3) + P(4)}{2}, \quad P_F = \frac{P(13) + P(14)}{2}. \tag{1}$$

To measure the overall temporal pitch variation within the profile, the pitch rising index, denoted by I_R, is defined as

$$I_R = k \cdot \frac{\mathrm{Max}_{i=2}^{15}\{P(i)\} - \mathrm{Min}_{i=2}^{15}\{P(i)\}}{\mathrm{Max}_{i=2}^{15}\{P(i)\} + \mathrm{Min}_{i=2}^{15}\{P(i)\}} \tag{2}$$

where

$$k = \begin{cases} 1 & \arg \mathrm{Max}_{i=2}^{15}\{P(i)\} > \arg \mathrm{Min}_{i=2}^{15}\{P(i)\} \\ -1 & \arg \mathrm{Max}_{i=2}^{15}\{P(i)\} \leq \arg \mathrm{Min}_{i=2}^{15}\{P(i)\} \end{cases}.$$

The first and the last segment of the pitch profile, i.e., $P(1)$ and $P(16)$, are not used in order to reduce possible errors in the detection of the voice portion. The polarity of I_R indicates the overall temporal trend of pitch movement within the utterance and the magnitude of I_R represents the degree of such variation.

B. Features that Separate Entering Tones from Nonentering Tones

It is noted that entering tones are much shorter in duration than nonentering tones since entering tone syllables must end with stop consonants /p/, /t/ or /k/. Therefore, the duration D of the voiced portion of a syllable can be used as a major discrimination factor to separate entering tones from nonentering tones. In normal speech, syllables of entering tones have much shorter voiced periods (typically 0.05–0.3 s) than those of nonentering tones (0.3–0.8 s). On the other hand, since the final stop in an entering tone syllable is glottalized, the utterance stops abruptly and the signal energy decreases very rapidly at the end of the syllable. This provides an additional feature to distinguish between entering and nonentering tones. To parametrize this characteristic, a smoothed short-time energy profile $\{E_S(j)\}$ is obtained for the voiced portion of the syllable [12]. Let $j_{\max}$ denote the maximum energy frame and t_d be the time required for energy to drop from 90% to 10% of $E_S(j_{\max})$. Then the energy drop rate is defined as the reciprocal of t_d, i.e.,

$$R_D = \frac{1}{t_d}. \tag{3}$$

Apparently, the duration D and the energy drop rate R_D are highly correlated. The shorter the duration is, the faster the energy drops. However, the rapid energy drop observed for an entering tone results mainly from the articulatory requirement of the final stop. This is because a syllable with a final stop consonant will stop sharply even if the voiced portion preceding the stop consonant has been prolonged. The classification hierarchy shown in Fig. 4 summarizes how to make use of these suprasegmental features in recognizing the lexical tone of Cantonese.

C. Normalization of the Feature Parameters

Normalization of the feature parameters is necessary because it will eliminate undesirable time and speaker variations of these parameters. Pitch is basically a physiologically determined characteristic and is regarded as being speaker dependent. To build a tone recognition system which uses the relative pitch level of input utterances as the main discriminative feature, a normalization procedure is needed to

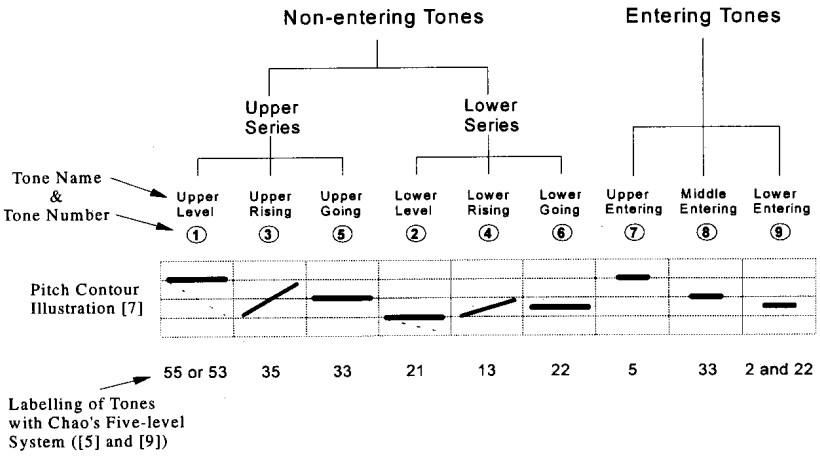

Fig. 3. Tones in the Cantonese dialect (dashed line means alternative pronunciation).

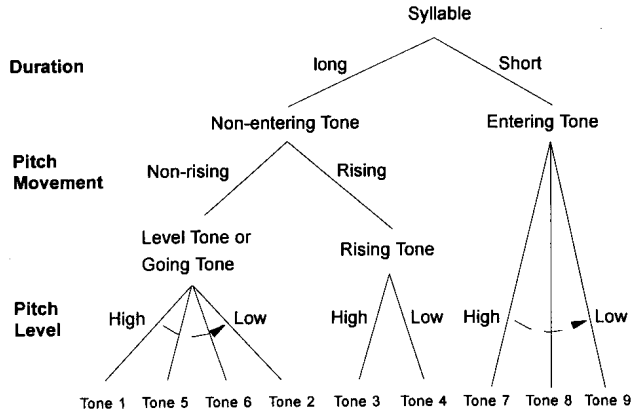

Fig. 4. The classification hierarchy for lexical tones of Cantonese.

align the range of pitch level for different speakers, thereby making the system performance speaker independent. Furthermore, it is also necessary to eliminate the undesirable effect due to the variation of the general key or "pitch of voice" of the same speaker over time [13]. The causes of such key changes may be emotional or physical. Sometimes intonational influences and stylistic factors may also affect the pitch level. The normalization of pitch is especially important for tone recognition of Cantonese. This is because there are so many distinct tones in Cantonese and their pitch levels are often very close to each other. As a result, confusion may occur if key change is not considered. The situation is even worse for male speakers since the dynamic pitch range of a male voice is much narrower (90–180 Hz) than that of a female voice (150–270 Hz).

For the duration-related parameters D and R_D, normalization is also needed. The speed of speaking can be affected by emotional, stylistic and environmental factors and it may change from time to time. The duration of nonentering tone syllables can be very short for fast speaking persons.

In our tests, normalization is done by simply dividing individual feature parameter by a factor, which is a representative value of that particular parameter over a certain period of time. Cheng [8] has found that the initial pitch values of syllables with tone 2, 3, 4, 6, and 9 have fairly small variance statistically. Therefore, we define the pitch normalization factor P_S as the mean of the initial pitch values of all the syllables with tone 2, 3, 4, and 6. Tone 9 is not included since the number of available syllables is very small. In fact, P_S can be regarded as an estimate of the intrinsic pitch of voice. In cases where syllables with tone 2, 3, 4, and 6 are not available, the initial pitch of other syllables can also be used but with a proper scaling. For example, the initial pitch of tone 1 is usually higher that of tone 2, 3, 4 or 6 and the scaling factor can be set to be 0.7.

The normalized initial pitch $\hat{P}_I$ and the normalized final pitch $\hat{P}_F$ are now given by

$$\hat{P}_I = \frac{P_I}{P_S}, \quad \hat{P}_F = \frac{P_F}{P_S}. \tag{4}$$

It should be noted that normalization is not required for the rising index of pitch I_R. The rising index is endowed by its definition (2) with a self-normalizing property and it is fairly insensitive to the shift of pitch range.

Similarly for parameters D and R_D, let the normalization factors be D_S and R_{DS} respectively. Then the normalized values, $\hat{D}$ and $\hat{R}_D$, can be computed as follows,

$$\hat{D} = \frac{D}{D_S}, \quad \hat{R}_D = \frac{R_D}{R_{DS}}. \tag{5}$$

In practice, D_S and R_{DS} are estimated from the same set of speech utterances which has been used to derive the pitch normalization factor as described earlier. Fig. 5 summarizes diagramatically the entire feature extraction process for tone recognition of Cantonese.

IV. The Neural Network Classifier

Artificial neural networks have been applied successfully to perform static pattern recognition. A three-layer feedforward network is employed in this tone recognition system as shown in Fig. 6. The input layer consists of five neurons, each representing a component of the extracted feature vector. To represent the nine different tones, the output layer is composed of nine separate units. A particular output unit will be activated to correspond to a particular tone. The size of hidden layer is task-dependent and is determined empirically. We have tried different number of hidden neurons in our tests, say 15, 20, 25, 30, 35, etc. and it was found that the best recognition performance was obtained by using 25 and 35 hidden neurons in single speaker and multispeaker applications respectively.

To test the performance of a properly trained classifier, the feature vector of an utterance is presented to the classifier and the output pattern is generated using current connection weights. If the ith output neuron exhibits the largest activation level (0.0–1.0), the utterance is recognized to carry a Cantonese syllable of tone i. To train a classifier, the features derived from the training utterances are first presented to the input units repetitively and the connection weights between neurons are then adjusted so as to reduce the error between the actual output patterns and the desired output patterns. In order to

TABLE II
THE DISTRIBUTION OF DIFFERENT TONES IN THE SPEECH CORPUS

Tone	Speech Corpus used in our study		Statistics by Fok [14]	
			15 minutes Radio Broadcast	45 minutes Conversation
1	34	14.53%	31.9%	28.6%
2	34	14.53%		
3	34	14.53%	28.8%	31.2%
4	33	14.10%		
5	32	13.67%	30.0%	31.0%
6	34	14.53%		
7	6	2.56%	4%	4.7%
8	11	4.70%	3.9%	1.4%
9	16	6.84%	1.2%	3.1%

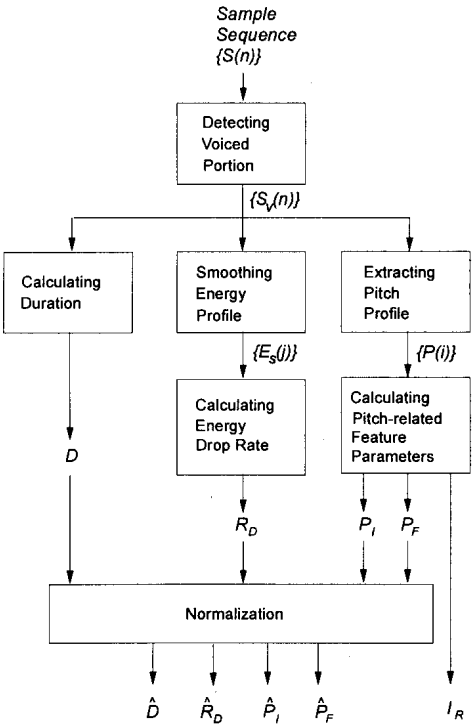

Fig. 5. Extraction of tone features.

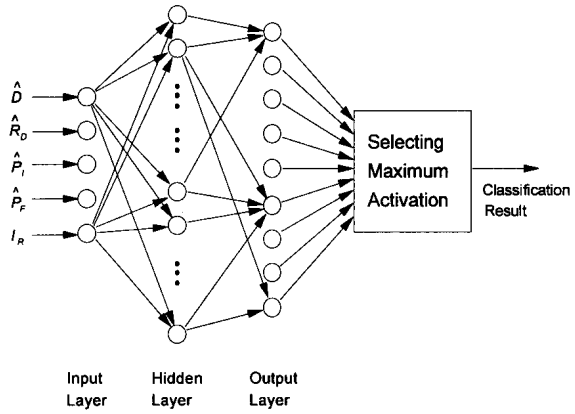

Fig. 6. The neural network classifier.

terminate a training process, the average output error of the training data is required to reach a preliminarily assigned threshold. The overall recognition accuracy for the training data is subsequently examined for each training cycle. If no further improvement on accuracy can be achieved after a certain time, the classifier is regarded as being properly trained and the training process will be stopped. The latter requirement is to prevent over-training of some of the data at the expense of the others, which may result in poor recognition performance.

V. PERFORMANCE EVALUATION OF THE PROPOSED SYSTEM

A. The Speech Corpus and Database

A speech corpus has been constructed for performance evaluation of the proposed tone recognition system. Since the Finals are regarded as prime carriers of distinctive pitch variation, the speech corpus covers all of the phonologically allowed combinations of 53 Cantonese Finals and 9 Cantonese tones, while the selection of Initials is arbitrary. The resultant speech corpus contains 234 Cantonese syllables. The distribution of tone 1–9 in the speech corpus is shown in Table II, along with the similar findings by Fok [14] for daily conversation and radio broadcasting speech. To make the speech corpus readable by the speakers, a commonly used Chinese character with the desirable Cantonese pronunciation is chosen for each of the 234 syllables. In a complete trial of recording, a subject is asked to read through the entire corpus, one character at a time. Ten subjects, 5 male and 5 female native Cantonese speakers, aged from 18 to 26, were invited to participate in preparing the speech database. Each speaker was asked to attend three trials of recording and three sets of data were recorded, each containing 234 utterances.

All speech data were recorded in a moderately quiet room with echo suppression. The speech signals were filtered by a low-pass filter with 4 kHz bandwidth and then digitized by a 14-bit A/D converter at 10 kHz sampling rate. An end-point detection algorithm was incorporated to cut out the leading and trailing nonspeech portions of the utterances. The utterances were screened manually and a small portion of them (about 2.3%) were discarded for their incorrect pronunciations or recording artifacts. The resulting database contained a total of 6857 utterances which were stored as binary files in a large capacity magnetic disk.

B. The Determination of Normalization Factors

A set of 24 syllables (about 10% of the corpus) were selected to obtain the normalization factors P_S, D_S, and R_{DS}. Since P_S is defined only by the initial pitch values of tone 2, 3, 4, and 6, this subset consists of syllables from these four tone classes only (six from each class). For both single speaker and multispeaker experiments, normalization of the tone features was performed for each set of speech data. The pitch period and the speed of speaking of utterances within a data set are assumed to be fairly constant but might be different from other data sets. The normalization factors

TABLE III

(a) CONFUSION MATRIX OF SINGLE-SPEAKER TONE RECOGNITION (TRAINING DATA). (b) CONFUSION MATRIX OF SINGLE-SPEAKER TONE RECOGNITION (TEST DATA)

	Recognized Tone										
Desired Tone	1	2	3	4	5	6	7	8	9	Total Count	Accuracy
1	1992	0	0	0	31	0	3	0	0	2026	98.3 %
2	0	1926	0	0	1	**46**	0	1	6	1980	97.3 %
3	0	0	1944	**41**	0	2	0	0	1	1988	97.8 %
4	0	4	22	1886	5	36	1	1	3	1958	96.3 %
5	32	1	0	2	1809	**49**	0	10	1	1904	95.0 %
6	1	**50**	1	4	38	1896	0	0	10	2000	94.8 %
7	0	0	0	0	0	0	329	1	0	330	99.7 %
8	0	0	0	0	5	0	1	593	**21**	620	95.6 %
9	0	3	0	1	0	3	0	**23**	878	908	96.7 %
										13714	96.6 %

(a)

	Recognized Tone										
Desired Tone	1	2	3	4	5	6	7	8	9	Total Count	Accuracy
1	955	1	1	0	**47**	0	5	4	0	1013	94.3 %
2	0	906	2	5	2	**63**	0	0	12	990	91.5 %
3	1	0	931	**57**	1	2	1	0	1	994	93.7 %
4	1	9	**48**	868	8	**36**	0	3	6	979	88.7 %
5	**52**	1	0	8	804	**73**	0	11	3	952	84.5 %
6	0	**49**	2	16	**55**	868	0	0	10	1000	86.8 %
7	2	0	0	0	0	0	149	**11**	3	165	90.3 %
8	6	0	2	5	**8**	3	3	238	**45**	310	76.8 %
9	0	**10**	3	3	0	**10**	2	**45**	381	454	83.9 %
										6857	89.0 %

(b)

TABLE IV

(a) CONFUSION MATRIX OF MULTISPEAKER TONE RECOGNITION (TRAINING DATA). (b) CONFUSION MATRIX OF MULTISPEAKER TONE RECOGNITION (TEST DATA).

	Recognized Tone										
Desired Tone	1	2	3	4	5	6	7	8	9	Total Count	Accuracy
1	1904	2	0	0	**98**	0	10	12	0	2026	94.0 %
2	0	1832	0	3	0	**121**	0	0	24	1980	92.5 %
3	2	1	1770	**210**	0	3	1	1	0	1988	89.0 %
4	0	6	35	1841	8	**59**	0	4	5	1958	94.0 %
5	**146**	1	1	11	1560	**155**	0	30	0	1904	81.9 %
6	0	**107**	2	14	**89**	1772	0	0	16	2000	88.6 %
7	3	0	0	1	0	0	302	**22**	2	330	91.5 %
8	6	1	1	4	**16**	0	**13**	513	**66**	620	82.7 %
9	0	9	0	9	4	**32**	1	**86**	767	908	84.5 %
										13714	89.4 %

(a)

	Recognized Tone										
Desired Tone	1	2	3	4	5	6	7	8	9	Total Count	Accuracy
1	940	1	1	2	**55**	0	5	9	0	1013	92.8 %
2	0	903	1	2	0	**64**	1	1	18	990	91.2 %
3	1	0	876	**115**	0	1	1	0	0	994	88.1 %
4	1	4	**22**	903	6	**37**	0	3	3	979	92.2 %
5	**85**	1	2	4	745	**98**	0	17	0	952	78.3 %
6	0	**53**	1	10	**49**	877	0	0	10	1000	87.7 %
7	1	0	1	0	0	0	147	**15**	1	165	89.1 %
8	5	1	4	3	**8**	1	**8**	239	**41**	310	77.1 %
9	0	7	1	6	1	**14**	0	**51**	374	454	82.4 %
										6857	87.6 %

(b)

were obtained by simply taking the average of the extracted initial pitch values, duration and energy drop rates of the 24 selected utterances. The feature vectors of all utterances in the same data set were normalized using these factors.

C. Experiments on Single-Speaker Tone Recognition

In the case of single-speaker tone recognition, both the training data and the test data came from the same subject. Three experiments were

performed for each of the 10 subjects by using utterances from one trial as the test data and the other two trials as training data. For all single-speaker experiments, 25 hidden units were used. The results are summed up as shown in Table III(a) and (b). Each highlighted entry indicates a contribution of over 2% misclassification for a particular tone.

The overall recognition rates for training and test data are given by 96.6% and 89% respectively. Tone 1 and 7, being characterized by their remarkably high pitch level, give the lowest error rates of recognition. Tone 3 also has a good recognition rate because of its distinctively sharp rise of pitch. Notable confusion is found within the lower nonentering tone series and between middle and lower entering tones since all of them have very close pitch levels as shown in Fig. 3. Similar reason also gives rise to classification errors between tone 1 and 5.

D. Experiments on Multispeaker Tone Recognition

Similar to the single-speaker case, the multispeaker tone recognition systems were trained with training data from two of the three trials of all subjects and tested with the remaining data sets. The combinations of training and test data were then permuted and two similar experiments were performed. The number of hidden units used was 35 for all multispeaker experiments. The results are shown in Table IV(a) and (b). The confusion patterns of multispeaker experiments are very similar to the single-speaker ones except for the increased number of recognition errors in individual entries. The overall recognition rates for training and test data are given by 89.4% and 87.6% respectively.

VI. Conclusion

In this paper an efficient method for the tone recognition of isolated Cantonese syllables has been proposed. This method utilizes the relative pitch levels, temporal pitch variation patterns and duration of voiced portion as the main discriminating features. The feature vector consists of five components and a three-layer feedfoward neural network is used for classification purpose. It has been shown by experimental results that the proposed method performs satisfactorily for both single-speaker and multispeaker recognition of a large vocabulary, and the average recognition accuracy is found to be 89.0% and 87.6% respectively.

Cantonese dialect is very rich in tones. Therefore tone identification plays an extremely important role in automatic recognition of Cantonese. If the lexical tone carried by an unknown syllable is correctly identified, the task of phonemes recognition will become much easier since the respective vocabulary size can be considerably reduced. The advantage of such a tone-oriented approach will be even more significant for connected speech recognition of Cantonese. This is because if the tone pattern of a polysyllabic utterance is known, the number of possible Cantonese phrases represented by that utterance can be minimized. Indeed, this process is commonly adopted in human perception of tonal speech.

References

[1] W. Yang, J. Lee, Y. Chang, and H. Wang, "Hidden Markov model for Mandarin lexical tone recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 7, July 1988.

[2] L. Liu, W. Yang, H. Wang, and Y. Chang, "Tone recognition of polysyllabic words in Mandarin speech," *Comput. Speech Language*, vol. 3, pp. 253–264, 1989.

[3] O.-K. Y. Hashimoto, *Studies in Yue Dialects 1: Phonology of Cantonese*. Cambridge, MA: Cambridge Univ. Press, 1972.

[4] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representation by error propagation," *Parallel Distributed Processing*, Vol. 1: Foundation. Cambridge, MA: MIT Press, 1986, ch. 8, pp. 318–364.

[5] J.-H. Yuan, "Hanyu fangyan gaiyao (the outline of Chinese dialects)," in *Wenzi Gaige Chubanshe*, 2nd ed., Beijing, 1983.

[6] "Yueh Yin Tu Li (a Cantonese syllabary with model pronunciations)," Edited by Ng Tor-Tai Chinese Language Research Centre, The Chinese Univ. of Hong Kong.

[7] S. L.Wong, *A Chinese Syllabary Pronounced According to the Dialect of Canton*. The Commercial Press, 1941.

[8] Y. H. Cheng, "An efficient classifier for speech recognition of Cantonese," MPhil. thesis, The Chinese Univ. of Hong Kong, 1991.

[9] Y.-R. Chao, *A Grammar of Spoken Chinese*, translated by Ting Pang Hsin, Chinese University Press, Hong Kong, 1980.

[10] A. Komatsu, A. Ichikawa, K. Nakata, Y. Asakawa, and H. Matsuzaka, "Phoneme recognition in continuous speech," in *Proc. ICASSP '82*, pp. 883–886.

[11] M. M. Sondhi, "New methods of pitch extraction," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, no. 2, pp. 262–266, June 1968.

[12] W. M. Lai, P. C. Ching, and Y. T. Chan, "Discrete word recognition using energy-time profiles," *Int. J. Electron.*, vol. 63, no. 6, pp. 857–865, 1987.

[13] K. L. Pike, *Tone Languages—A Technique for Determining the Number and Type of Pitch Contrasts in a Language, with Studies in Tonemic Substitution and Fusion*. Ann Arbor, MI: The Univ. of Michigan Press, 1948.

[14] Y.-Y. Fok, "A perceptual study of tones in Cantonese," Centre of Asian Studies, Univ. of Hong Kong, 1974.

Adaptive WRLS-VFF for Speech Analysis

D. G. Childers, J. C. Principe, and Y. T. Ting

***Abstract*— The purpose of this correspondence is to show that an adaptive weighted recursive least squares algorithm with a variable forgetting factor (WRLS-VFF) will adjust the size of the data segment to be analyzed according to its time-varying characteristics, as during the transitions between vowels and consonants. The algorithm can accurately estimate the vocal tract formants, anti-formants, and their bandwidths, be used for glottal inverse filtering, perform voiced (V)/unvoiced (U)/silent (S) classification of speech segments, estimate the input excitation (either white noise or periodic pulse trains), and estimate the instant of glottal closure.**

I. Algorithm Description

We assume that the speech signal is generated by an autoregressive, moving average (ARMA) model

$$y_k = -\sum_{i=1}^{p} a_i(k) y_{k-i} + \sum_{j=1}^{q} b_j(k) u_{k-j} + u_k \qquad (1)$$

Manuscript received August 13, 1993; revised November 15, 1994. This work was supported by NIH Grant NIDCD R01 DC00577, NSF Grant IRI-9215331, the University of Florida Center of Excellence Program in Information Transfer and Processing, and the Mind Machine Interaction Research Center. The associate editor coordinating the review of this paper and approving it for publication was Dr. Amro El-Jaroudi.

D. G. Childers and J. C. Principe are with the Department of Electrical Engineering, University of Florida, Gainesville, FL 32611 USA.

Y. T. Ting is with the Chung San Institute of Science and Technology, Taiwan, Republic of China.

IEEE Log Number 9410229.