Proceedings of 1993 International Joint Conference on Neural Networks

An NN Based Tone Classifier for Cantonese

Tan LEE*, P.C. CHING*, L.W. CHAN and Brian MAK***

* Department of Electronic Engineering
The Chinese University of Hong Kong
Shatin, Hong Kong

** Department of Computer Science
The Chinese University of Hong Kong
Shatin, Hong Kong

ABSTRACT

Chinese language is a typical monosyllabic tonal language. Tone identification is undoubtedly an essential component in the speech recognition problem of Chinese, specifically for the Cantonese dialect which is well known of being very rich in tones. This paper presents an efficient method for tone classification of isolated Cantonese syllables. Several suprasegmental feature parameters for tone identification are extracted from the voiced portion of each recorded utterance and then fed into a multi-layer neural network classifier. Using a large vocabulary containing 234 distinct syllables, the system performance for single-speaker and multi-speaker cases are found to be 89% and 87% respectively.

1 INTRODUCTION

Cantonese is one of the most commonly used dialects in the Southern China region, and it is the mother tongue of forty millions people all around the world. In Hong Kong, most of the residents speak Cantonese in their daily communication. Like most of the other Chinese dialects, Cantonese is monosyllabic and tonal. A Chinese character, when being pronounced in Cantonese, is uttered as a single syllable word which has an intrinsic lexical tone pattern. To recognize a Cantonese syllable, not only its phonetic transcription but also its lexical tone have to be identified.

In this paper a tone classification system for isolated Cantonese words is proposed (Figure 1). By isolated word we mean that each of the input utterances is composed of a single Cantonese syllable sound. The proposed approach uses the pitch profile of individual utterance as a basis upon which suprasegmental feature information is extracted. In order to minimize the computational complexity of the classification system, only five parameters are derived for an utterance to signify its distinctive tone features. These parameters are then fed into a neural network classifier which presents the classification results at its output neurons. A three-layer perceptron is adopted and training of the neural network is carried out using the standard error backpropagation algorithm[7]. For simulation in single-speaker and multi-speaker experiments, the number of hidden units used are 25 and 35 respectively.

In the next section a brief introduction on the phonology of Cantonese will be given and the acoustic features of the nine Cantonese tone patterns will be identified. In Section 3 we will explain in detail how to obtain a set of suprasegmental feature parameters from the recorded speech samples. Then the experiments that have been performed for performance evaluation will be described and the attained recognition results are given.

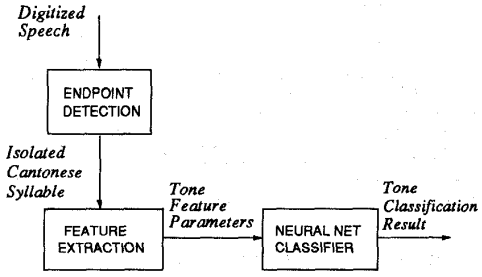

Figure 1 : The Proposed Tone Classification System for Isolated Cantonese Syllables

2 TONE SYSTEM OF CANTONESE

A Cantonese syllable can be simply divided into an *Initial* part and a *Final* part[1]. The *Initial* part is optional and if exists, is either an unvoiced or a nasal consonant. The *Final* part usually consists of the concatenation of a middle vowel and an optional final consonant which is either a voiced or a stop consonant.

There are totally 19 *Initials* and 53 *Finals* in Cantonese, making up more than 1500 different phonetic transcriptions and about 700 of them are used in the dialect. Furthermore there are nine different tones and each of the 700 phonetic transcriptions may appear in more than one distinct syllable which differ from each other only by their lexical tones. As a result, a total of about 1,600 valid syllable sounds are found in contemporary Cantonese dialect[1].

The nine Cantonese lexical tones can be categorized into two groups: the *entering* and *non-entering* tone groups. Traditional phonology of Cantonese divides the six *non-entering* tones into upper and lower series, according to their relative pitch levels. Each of the upper and lower tone series is composed of a *level* tone, a *rising* tone and a *going* tone. A syllable in the *entering* tone group must be ended with a stop consonant (p,t or k) and thereby can be characterized by a very short duration of the corresponding speech utterance. Moreover, it is also noted that a speech utterance of *entering* tone exhibits a rapid amplitude drop at the end portion.

Figure 2 depicts the nine different tone patterns of Cantonese by showing their relative pitch levels, pitch movement and relative durations. The nine tones are usually referred as tone 1 to tone 9 as shown in Figure 2. From this illustration we note that,

i) each of the *non-entering* tones can be distinguished from the others either by its relative pitch level or by the temporal variation of its pitch level.

ii) the so-called *entering* tones do not have their own patterns of pitch movement. The pitch values of the upper, middle and lower *entering* tones are identical with that of the *upper level*, *upper going* and *lower going* tones respectively.

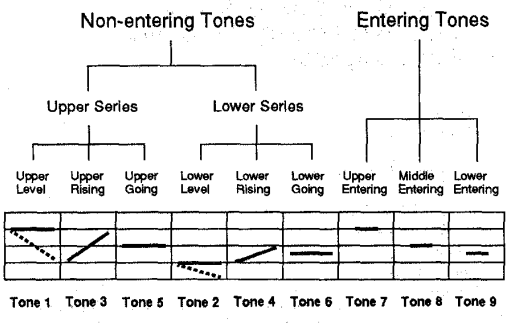

Figure 2 : The Nine Tone Patterns of Cantonese

Based on the above observations, three suprasegmental features, namely the duration of the utterance, the relative pitch level and the rising rate of pitch value within the utterance, have been selected for the purpose of tone identification in our study.

3 SUPRASEGMENTAL FEATURE PARAMETERS FOR TONE CLASSIFICATION

The feature extraction process described below can be summarized as in Figure 3.

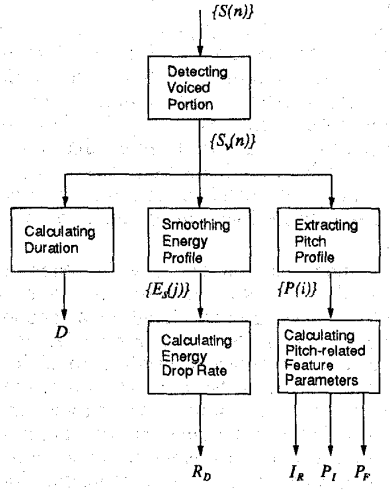

Figure 3 : Extraction of Suprasegmental Feature Parameters

Detection of Voiced Portion

For an isolated Cantonese syllable, the tone information is carried by the voiced portion of the utterance. Therefore, the first step of extracting tonal features is to locate the voiced segment of an utterance accurately. All feature parameters will be derived from the speech samples within the voiced portion.

Lai *et al*[2] suggested that segmental energy and zero-crossing rate can be used as the major discriminating features to extract the voiced portion of a Cantonese syllable. A revised version of this approach has been adopted in the present tone classification system.

The input speech sequence $\{S(n)\}$ is first divided into 10 ms time frames with 50% overlapping. For each of the 10 ms frames, the frame energy and zero-crossing rate are computed, resulting two profiles $\{E(j)\}$ and $\{Z(j)\}$, where $E(j)$ and $Z(j)$ denote the energy and zero-crossing rate of the jth frame respectively. The beginning and the end of voiced portion can be determined by searching backward and forward respectively from the frame with maximum frame energy. The beginning point is assumed when the energy of a particular frame is smaller than a threshold E_{it} or the zero-crossing rate is higher than a threshold Z_t. For locating the end point, only an energy threshold E_{ft} is employed. By using this method with properly selected thresholds, the voiced portion of a Cantonese syllable can be detected accurately in most cases.

Pitch-related Feature Parameters

To obtain the pitch profile of an utterance, the detected voiced speech is divided into 16 frames with equal duration. The pitch value of each frame is calculated from the speech samples within that frame. This warping process linearly aligns speech sequences with different durations and it has been shown to perform quite satisfactorily for small variations in speech length[3].

The pitch extraction method adopted here is a modified version of the 3-level center clipped pitch extraction algorithm proposed by Sondhi[4]. The resulted pitch profile is represented by a vector $\{P(1),P(2),\ldots,P(i),\ldots,P(16)\}$ where $P(i)$ denotes the pitch value of the ith aligned frame. Based on the time-aligned pitch profiles, the *initial pitch* P_I and the *final pitch* P_F of the speech utterance are defined as follows,

$$P_I = (P(3)+P(4))/2 \quad \ldots (1)$$

$$P_F = (P(13)+P(14))/2 \quad \ldots (2)$$

Another pitch-related parameter I_R, referred as the *rising index*, is calculated from $\{P(i)\}$ as,

$$I_R = k\frac{\text{Max}\{P(i)\}-\text{Min}\{P(i)\}}{\text{Max}\{P(i)\}+\text{Min}\{P(i)\}}, \qquad 3 \le i \le 14$$

where

$$k = \begin{cases} 1 & arg\ \text{Max}\{P(i)\} > arg\ \text{Min}\{P(i)\} \\ -1 & arg\ \text{Max}\{P(i)\} \le arg\ \text{Min}\{P(i)\} \end{cases} \quad \ldots (3)$$

The polarity of I_R indicates whether the pitch level within an utterance rises or drops with time while the magnitude of I_R measures the degree of such rise or drop. This feature parameter has a self-normalizing property and it is fairly insensitive to the speaker-dependent pitch ranges.

Duration and Energy Drop Rate

The *duration D* of an isolated Cantonese word is defined as the length of the voiced portion of the utterance. If the detected voiced segment contains N_v consecutive 10 ms frames (with 50% overlapping), D can be evaluated as,

$$D = (N_v + 1) \times 5 \quad msec. \qquad \ldots (4)$$

To calculating the rate of energy drop at the end of a discrete utterance, the energy profile $\{E_v(j)\}$ of the voiced portion is first smoothed using moving average approach with a window length of 5 frames. The smoothed energy profile $\{E_{sv}(j)\}$ is computed as,

$$E_{sv}(j) = \frac{1}{5} \sum_{m=-2}^{2} E_v(j+m) \qquad \ldots (5)$$

Let j_{max} denote the frame with maximum smoothed frame energy and t_d be the time required for $\{E_{sv}(j)\}$ to decline from 90% to 10% of $E_{sv}(j_{max})$. Then the *energy drop rate* R_D is defined by

$$R_D = 1/t_d \qquad \ldots (6)$$

Normalization of the Feature Parameters

It is obvious that the pitch-related parameters derived above are highly speaker-dependent. Speakers at different ages certainly exhibit very diversified dynamic ranges of pitch. Female speakers usually have higher pitch levels than male. Moreover, the pitch range of the same speaker may also vary to some extent from time to time. As a result, those tones with very close pitch levels, e.g. tone 2, tone 5 and tone 6, are easily confused with each other even for the same speaker. Similar problems also exist for the duration-related parameters D and R_D. The durations of non-entering tone syllables can be very short for those people who are used to speaking fast. Therefore a normalization procedure is needed to deal with the confusion problems caused by speaker dependence as well as temporal fluctuation of feature parameters.

The normalization procedure for an individual speaker is to divide each of the speaker-dependent feature parameters by a properly estimated normalization parameter. That is, for the *initial pitch* and the *final pitch*, we have,

$$\hat{P}_I = P_I / P_S \qquad \ldots (7)$$

$$\hat{P}_F = P_F / P_S \qquad \ldots (8)$$

where P_S is referred as the *intrinsic pitch* value of this particular speaker, and is defined as the mean of initial pitch values of all utterances in tone 2,3,4 and 6.

While the *normalized duration* $\hat{D}$ and the *normalized energy drop rate* $\hat{R}_D$ are given respectively by,

$$\hat{D} = D / D_S \qquad \ldots (9)$$

$$\hat{R}_D = R_D / R_{SD} \qquad \ldots (10)$$

The normalization parameters D_S and R_{SD} of each individual speaker are the average duration and rate of energy drop evaluated from a pre-selected vocabulary.

4 NEURAL NET CLASSIFIER

The neural network classifier is composed of a three-layer perceptron followed by a winner-take-all decision unit. The perceptron network consists of five input units for the five feature parameters as defined above, and nine output units, each representing a lexical tone class. In the cases of single-speaker and multi-speaker tone classification, the number of hidden units are 25 and 35 respectively. The classification outcome is indicated by the output unit with the maximum activation.

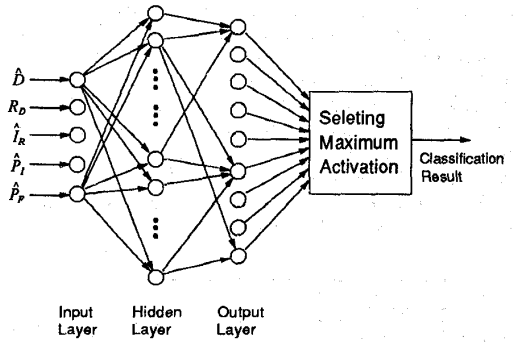

Figure 4 : The Neural Network Classifier for Tone Classification

5 SPEECH CORPUS

A speech corpus for performance evaluation of the proposed approach has been designed very carefully. Since the *Finals* have been regarded as the prime carriers of distinctive pitch variation[6], the speech corpus should cover all the phonologically allowed combinations of the 53 Finals and the 9 tones, while the selection of Initials is arbitrary.

As a result, a total of 234 valid syllable sounds are collected in the speech corpus. Each of the syllables may correspond to a number of Chinese characters among which a commonly used one is selected as the representative. The speech corpus is therefore a set of 234 single syllable Chinese words. To prepare a full set of training or testing data, each object speaker is asked to read the complete set of characters. The recorded utterances are then stored in separate files for further processing.

6 SIMULATION RESULTS

All recording work has been done in a 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 an trailing non-speech segments of an utterance, resulting in a speech sample sequence ready for tone feature extraction.

A total of 10 native Cantonese speakers, 5 male and 5 female, aged from 20 to 25, were invited to contribute to the speech database. Each individual speaker was asked to attend three trials of recording, each producing a set of 234 utterances according to the speech corpus. A recording trial was required strictly to be conducted continuously in order to make the fluctuation of dynamic pitch range as small as possible.

A set of 24 Cantonese syllables were selected from the speech corpus for generating the normalization parameters P_S, D_S and R_{SD}. For a particular trial P_S, D_S and R_{SD} were computed by averaging the extracted initial pitch values, durations and energy drop rates of the 24 representative utterances obtained during this trial. Then the normalization was done for all utterances in the same trial according to Eqn.(7) - (10).

In the simulation of a single-speaker tone classification system, both the training data and testing data come from the same speaker. For each of the ten speakers, the speech data obtained in trial 1 and trial 2 were used for neural network training and those in trial 3 are used as test data. In order to provide a thorough evaluation, we have shuffled the grouping of training data and test data among the three trials and repeated the experiment. The resulted confusion matrices are shown in Table 1 (a) & (b) and the overall recognition rates for training data and test data are 96.6 % and 89.0 % respectively.

	Recognized As Tone										
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*Total*	*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 %

Table 1(a): Confusion Matrix of the **Single Speaker** Tone Classification System (Training Data)

	Recognized As Tone										
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*Total*	*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 %

Table 1(b): Confusion Matrix of the **Single Speaker** Tone Classification System (Test Data)

The training data in the multi-speaker case employs the first two trials of all the 10 speakers and the corresponding test data adopt all trial 3 data. Similar to the single-speaker case, a shuffling procedure is incorporated to obtain another two different combinations of training data and test data. The overall recognition rates for training data and test data are 89.4 % and 87.6 % respectively and the confusion matrices are shown in Table 2(a) & (b).

	Recognized As Tone										
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*Total*	*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 %

Table 2(a): Confusion Matrix of the **Multi-Speaker** Tone Classification System (Training Data)

	Recognized As Tone										
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*Total*	*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 %

Table 2(b): Confusion Matrix of the **Multi-Speaker** Tone Classification System (Test Data)

In this paper we have introduced an efficient and effective tone classification scheme for Cantonese dialect using speech processing techniques and artificial neural networks. Preliminary simulation results show that the acoustic features for tone identification have been correctly utilized and the neural network classifier performs satisfactorily in this application.

7 REFERENCES

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

[2] W.M. Lai, P.C. Ching and Y.T. Chan, *Discrete Word Recognition Using Energy-Time Profiles*, Int. J. Electronics 63, No.6, pp.857-865, 1987.

[3] A. Komatsu, A. Ichikawa, K. Nakata, Y. Asakawa and H. Matsuzaka, *Phoneme Recognition in Continuous Speech*, Proceedings of ICASSP, pp.883-886, 1982.

[4] M.M. Sondi, *New Methods of Pitch Extraction*, IEEE Trans. on Audio and Electroacoustics, AU-16, No.2, pp.262-266, 1968.

[5] Y.H. Cheng, *An Efficient Tone Classifier for Speech Recognition of Cantonese*, MPhil. Thesis, The Chinese University of Hong Kong, 1991.

[6] M. Delos, B. Guerin, M. Mrayati and R. Carre, *Study of Intrinsic Pitch of Vowels*, J. Acoustic Soc. Am. 59, Suppl. No. 1, pp.572, 1976.

[7] D.E. Rumelhart, G.E. Hinton and R.J. Williams, *Learning Internal Representation by Error Propagation*, Parallel Distributed Processing, Volume 1: Foundation, pp.318-364, MIT Press, 1986.