Automatic Recognition of Continuous Cantonese Speech with Very Large Vocabulary

Ying Pang Alfred NG[1], L. W. CHAN[1], P. C. CHING[2]
[1]Department of Computer Science and Engineering [2]Department of Electronic Engineering
The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong
Tel: (852) 2609 8411 Fax: (852) 2603 5302
Email: ypng@cs.cuhk.edu.hk

Abstract

This paper presents the first published results for automatic recognition of continuous Cantonese speech with very large vocabulary. The size of the vocabulary covered by this system is about the same as that encountered in the Hong Kong local Chinese newspaper, Wen Hui Bao (文匯報). The system covers 6335 Chinese characters (字) and a large number of Chinese words (詞) can be formed by combining these Chinese characters. The input to the system is the end pointed speech waveform of a sentence or phrase, the output is the Big5 coded Chinese characters. In the development of the recognition system, we have devised new methods in 1) construction of a continuous Cantonese speech database, 2) lexical tone recognition in continuous Cantonese speech, and 3) integration of lexical tone and base syllable recognition results. The speaker dependent recognition rates for Chinese character, base syllable and lexical tone are 90.94%, 94.73% and 69.7% respectively.

Introduction

Automatic speech recognition can be applied in many areas e.g. dictation machine, interactive voice enquiry system using telephone and aid for hearing impaired etc. One area which speech recognition is highly desired is Chinese character input via speech. It is very difficult to enter Chinese characters at a speed comparable with typing in English words. There are different Chinese input platforms. The first one is by typing on a keyboard and there are different input methods e.g. CangJie (倉頡), Array (行列), Simple (簡易) and PinYin (拼音) etc. All of them demand a certain extent of extra training and the input speed depends on the input method's coding effectiveness and the user's familiarity with the method. These input methods are cryptic for non-trained people. The second one is by hand written character recognition. However, the user must know how to write Chinese characters. Besides, the input speed is quite slow (typically 12 Chinese characters per minute). A much faster and natural Chinese character input means can be achieved if real time Chinese (e.g. Mandarin or Cantonese) speech recognition is feasible.

Automatic speech recognition work is performed on Cantonese (rather than some other Chinese dialects e.g. Mandarin) because of various reasons. Firstly, Cantonese (廣東話 or 粵語) is a popular Chinese dialect which is spoken by tens of millions of people living in Hong Kong and the southern part of Mainland China. Secondly, Cantonese is a challenging language for research. Cantonese has a complicated tone system. It has 6 or 9 lexical tones. Very little work has been performed on automatic recognition of Cantonese speech. Only a few papers have been published on isolated Cantonese syllable recognition [1] and Cantonese lexical tone recognition in isolated syllables [2].

In Cantonese Chinese, there are at least 60,000 commonly used Chinese words (詞). Each of the words is composed of one to several Chinese characters (字). There are about 6335 commonly used Chinese characters and each of them is monosyllabic. Totally, there are about 1400 phonologically allowed Cantonese syllables [3]. Cantonese is a tone language and each syllable (e.g. /si2/) can be divided into two parts: the lexical tone (e.g. tone 2) and the base syllable (e.g. /si/). In Cantonese, there are six lexical tones [4] which are mainly characterised by the relative fundamental frequency (Fo) contours as shown in figure 1.

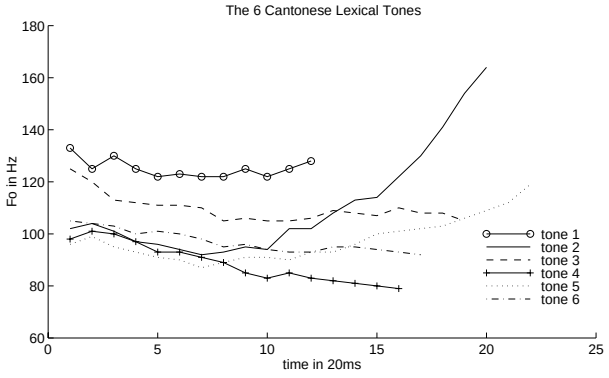

Figure 1: The 6 Cantonese lexical tones produced by a male speaker

There are about 555 Cantonese base syllables. In automatic Cantonese speech recognition, both lexical

tone and base syllable have to be recognised. It can be assumed that tone realisation and base syllable realisation are independent of each other. Therefore, they are recognised separately. The functional block diagram of our continuous Cantonese speech recognition system is shown in figure 2. Continuous Cantonese

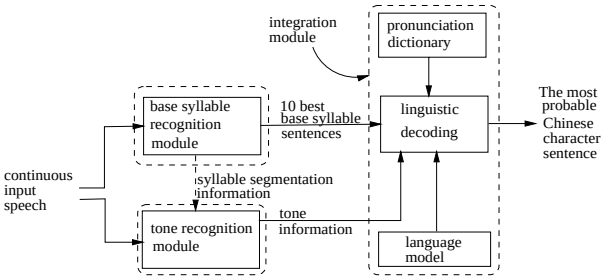

Figure 2: Functional block diagram of the continuous Cantonese speech recognition system

speech in the form of a sentence or a phrase is processed to obtain some speech parameters (e.g. mel frequency cepstral coefficients). The speech parameters are input into the base syllable recognition module to obtain 10 best base syllable sentences. At the same time, continuous speech is also processed to give speech parameters (e.g. fundamental frequency) for the tone recognition module. The continuous speech stream is segmented into syllables. This segmentation is performed by the base syllable recognition module. For each syllable, the probabilities for all possible Cantonese lexical tones are given by the tone recognition module. The tone information and the 10 best base syllable sentences are fed into the integration module. Base syllable is used to access Chinese characters by means of a Cantonese pronunciation dictionary. The tone probabilities and language model are used to find the most probable Chinese character sentence for the input speech.

Continuous Cantonese Speech Database Establishment

A continuous Cantonese speech database is crucial in the development of a continuous speech recognition system as the database is needed in the training and testing of the recognition system. Till now, there is no published continuous Cantonese speech database for automatic speech recognition. Hence, we have to build our own continuous Cantonese database. Before the sound recording of a speech database can be performed, a set of appropriate sentences need to be compiled. A phonetically balanced set of sentences is preferred because such a set has a number of advantages. Firstly, a speech recogniser can be trained using a minimal amount of speech data. Secondly, speech recognition module and language model can be trained separately to deal with different speech recognition tasks. In the past, the set of sentences was composed by a group of experts (e.g. phoneticians or specially trained linguists). The compilation process was difficult and labour intensive. In our project, we have applied a new, novel and semi-automatic approach to provide the set of sentences. The 2400 training sentences and 1200 testing sentences are selected from a newspaper text database which has about 2.2 million sentences.

Firstly, a Chinese text database is collected and transcribed automatically. Newspaper passages are provided by a Hong Kong local Chinese newspaper (Wen Hui Bao 文匯報). Sentences (or phrases) are segmented from the passages mainly by punctuation marks. The sentences are automatically transcribed by looking up a Cantonese pronunciation dictionary. This transcribed text database has about 29 million Chinese characters which make up about 2.2 million sentences or phrases.

Secondly, 3600 sentences are selected according to the phone balance criterion. One sentence is selected from the transcribed Chinese text database at a time. Each sentence in the text database is given a score in respect of the phone distribution after adding the sentence to the current speech database. The sentence to be selected is the one which has the highest score. After the sentence is added to the speech database, the whole process of sentence scoring and selection will be repeated until the required number of sentences are obtained for the speech database. Two procedures are used to maintain phone balance:-

A) Sequential selection of rare phonemes

It is found that the phonemes /m_s, kw, ng_s/ are relatively rare in the transcribed text database. Hence, a certain number of sentences containing the rare phonemes are specifically selected first.

B) To use a phone scoring method to maintain phone balance in the course of selection

The Variance method for phone scoring is adopted.

$$phone\_score = \frac{\sum_{i=1}^{n}(freq_i - \bar{x})^2}{\bar{x}}$$

In the equation, n is all 43 phonemes which are used in the speech recogniser, $freq_i$ is the frequency of the phoneme i for all the sentences chosen so far after a certain sentence is selected, and $\bar{x}$ is the average frequency of all the phonemes.

Thirdly, the 3600 selected sentences are manually rectified for the errors caused by automatic phonemic transcription.

The phone distribution of both the 2400 training sentences and the 1200 testing sentences are quite close to an even distribution. For instance, the 2400 training sentences have phone distribution in the interval of [0.940%, 3.242%] with a standard deviation of 0.633%[1]. Thus, the proposed semi-automatic selec-

[1]For an ideal even distribution, all the phones should have a percentage of 2.326%. The phone distribution of the original 2.2 million sentence text database is in the range of [0.002%, 8.970%] with a standard deviation of 1.955%.

tion approach can actually select a nearly phonetically balanced set of sentences.

A male native speaker of Cantonese was requested to read aloud for all 3600 sentences. The recording was performed in an ordinary laboratory using the Gradient Desklab 216 and a close talking microphone (Shure SM10A). The sampling rate was 16 KHz and each sample was coded by a signed 16 bit short number.

Lexical Tone Recognition in Continuous Cantonese Speech

Lexical tone is an integral part of a tone language and Cantonese is well known of being very rich in tone. The lexical tone of a syllable is mainly depicted by the syllable's fundamental frequency (Fo) contour. When compared with tone recognition in isolated syllable, tone recognition in continuous speech is more difficult. Firstly, the Fo contour has to be segmented from the continuous speech input stream. Secondly, the Fo contour is different in different sentential positions. Thirdly, the Fo contour is affected by its immediate tone context. The speech segmentation problem is tackled by the Hidden Markov model (HMM) base syllable recogniser to be mentioned in the next section. We have devised a new and efficient method to solve the sentential position problem and the immediate tone context problem [5]. The sentential position information and the immediate Fo context information are added as extra input nodes in a multilayer perceptron (MLP). It is found that the MLP can perform the Fo normalisation automatically. The adopted tone MLP configuration is 32-20-6[†]. The 32 input nodes include the current Fo contour (10 input nodes), preceding Fo contour (10 nodes), following Fo contour (10 nodes) and sentential position information (2 nodes) [5]. This approach is compact and efficient as only one single MLP is needed. We have achieved a speaker dependent tone recognition rate of 69.7% for testing data of continuous Cantonese speech.

Base Syllable Recognition

A Cantonese base syllable has a general structure of C_1VC_2 where C_1 is a syllable initial consonant, V is a vowel and C_2 is a syllable final consonant. There are 20 syllable initial consonants, 20 vowels and 7 syllable final consonants. The HMM approach is used for base syllable recognition in continuous Cantonese speech[2]. The units of recognition are 87,077 cross base syllable triphones as they can account for the coarticulation effect which occurs in continuous Cantonese speech. By a process of decision tree based clustering [6], they are reduced to 3478 tied state triphones. The recognition network with base syllables as its components is depicted in figure 3. Each base syllable composite HMM (e.g. /wut/ HMM) is formed by concatenating the appropriate clustered cross base syllable triphone HMM's (e.g. /w/ HMM + /u/ HMM + /t_f/ HMM). Each tied state triphone is represented by a

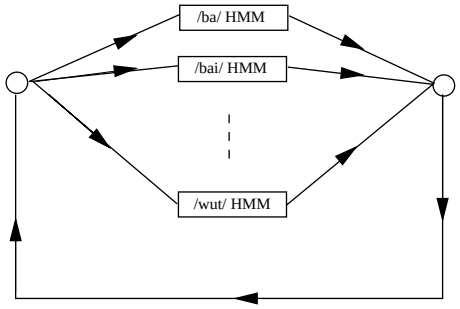

Figure 3: The base syllable HMM recognition network

continuous density HMM which has 3 emitting states and each state has 6 Gaussian mixtures. The speech parameters are 12 mel frequency cepstral coefficients and one normalised energy. In addition, the first order and second order time derivatives of the 13 coefficients are used. The speech waveform frame size is 25 ms and the frame period is 10 ms.

In the HMM base syllable recognition network, we have applied a base syllable backoff bigram to control the transition from one base syllable to another base syllable. This base syllable bigram improves the performance of the base syllable recogniser significantly[3]. The speaker dependent base syllable recognition rates using top 1 and top 10 base syllable sentences are 94.73% and 98.47% respectively for continuous Cantonese speech.

Integration of Base Syllable and Tone Recognition Result

We have devised a new and efficient method to integrate the base syllable and lexical tone recognition results in the search for the most probable Chinese character sentence. Initially, tone enriched 10 best sentences are found. The 6 lexical tone scores are provided by the MLP tone recognition module. The base syllable 10 best sentences are given by the HMM base syllable recognition module. Each base syllable sentence is composed of many base syllables. Each base syllable may correspond to many Chinese characters according to the pronunciation dictionary with multiple pronunciations. Then, a Chinese character lattice is formed by each base syllable sentence. An example of such a lattice is shown in figure 4. In the figure, 'ch' stands for a Chinese character, 't' is the tone score for the corresponding Chinese character

[†]32-20-6 means 32 input nodes, 20 hidden nodes and 6 output nodes. Nodes between adjacent layers are fully connected.

[2]The system prototype is implemented via HMM Toolkit V2.0.

[3]For example, using top 1 base syllable sentence, the base syllable recognition rate is raised from 90.66% to 94.73%.

e.g. '$t1_m$' is the tone score for the Chinese character '$ch1_m$'. In the lattice, a dynamic programming search

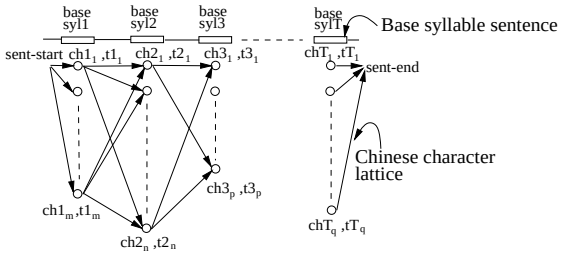

Figure 4: An example of Chinese character lattice

algorithm [7] is used to search for the best Chinese character sentence representing the base syllable sentence. Each transition's probability (shown as arrow in the Chinese character lattice) is obtained from the backoff Chinese character bigram calculated by using the 2.2 million sentences in the text database.

Assume that there is a transition from a Chinese character, ch_i, into a certain Chinese character, ch_j as shown in figure 5. $trans_i$ is the transition probability from ch_i into ch_j and it is obtained from the backoff Chinese character bigram. t is the lexical tone score for the Chinese character ch_j.

$$\mathrm{ch_i} \xrightarrow{\mathrm{trans_i}} \mathrm{ch_j},\ \mathrm{t}$$

Figure 5: Transition from Chinese character ch_i into Chinese character ch_j

The log probability for transition into the Chinese character, ch_j, is $\log(prob)$:-

$$\log(prob) = w_1 \times \log(trans_i) + w_2 \times \log(t)$$

where w_1, w_2 are weighting factors.

After the search in the Chinese character lattice has finished, the best Chinese character sentence would have a weighted log language and log tone score, *combine*. This is obtained by tracing back along the optimal path in the Chinese character lattice and summing the log probability for each transition. Previously, from the HMM base syllable recognition system, the base syllable sentence has obtained a log acoustic score. The two scores are combined to give a composite score as follows:-

$$\text{composite score} = \log(\text{acoustic score}) + w \times combine$$

where w is a weighting factor.

As top 10 base syllable sentences are used, accordingly 10 Chinese character sentences are found. The 10 Chinese character sentences will be reranked by their composite scores. The most probable Chinese character sentence is the one with the highest composite score. For Wen Hui Bao's 1200 testing sentences, the speaker dependent Chinese character recognition rate is 90.94%. Significant improvement in Chinese character recognition result is achieved by integrating the lexical tone and base syllable recognition results[4]. The performance of the system is further verified by selecting 50 sentences from another local Chinese newspaper, Ming Pao (明報) and the Chinese character recognition rate is 89.76% which is about the same as that for Wen Hui Bao.

Conclusions

We have developed the first published automatic speech recognition system of continuous Cantonese speech. In the development of the system, we have devised new methods in 1) continuous Cantonese speech database establishment, 2) lexical tone recognition in continuous Cantonese speech, and 3) integration of lexical tone and base syllable recognition results. Moreover, by using a base syllable backoff bigram in the HMM base syllable recognition network, the base syllable recognition results have been improved significantly. In our system, the speaker dependent recognition rates for Chinese character, base syllable and lexical tone are 90.94%, 94.73% and 69.7% respectively.

References

[1] Tan Lee, P. C. Ching, and L. W. Chan. An RNN Based Speech Recognition System with Discriminative Training. In *EUROSPEECH'95*, volume 3, pages 1667–1670, 1995.

[2] Tan Lee, P. C. Ching, L. W. Chan, Y. H. Cheng, and Brian Mak. Tone Recognition of Isolated Cantonese Syllables. *IEEE Transactions on Speech and Audio Processing*, 3(3):204–209, May 1995.

[3] Sek Ling Wong. *A Chinese Syllabary Pronounced According to the Dialect of Canton.* Chung Wah Bookstore, 1941.

[4] Daniel Jones and K. T. Woo. A Cantonese Phonetic Reader, 1912.

[5] Alfred Ying Pang Ng, P. C. Ching, and L. W. Chan. Automatic Recognition of Cantonese Lexical Tones in Connected Speech by Multi-Layer Perceptron. In *EUROSPEECH'95*, volume 3, pages 2205–2208, Madrid, September 1995.

[6] P. C. Woodland, J. J. Odell, V. Valtchev, and S. J. Young. Large Vocabulary Continuous Speech Recognition Using HTK. In *ICASSP'94*, volume II, pages 125–128, 1994.

[7] J. N. Holmes. *Speech Synthesis and Recognition.* Van Nostrand Reinhold (UK) Co. Ltd., 1988.

[4] The Chinese character recognition rate by using base syllable recognition results alone is 87.21%.