A NEURAL NETWORK BASED SPEECH RECOGNITION SYSTEM FOR ISOLATED CANTONESE SYLLABLES

Tan Lee and P.C. Ching
Department of Electronic Engineering
The Chinese University of Hong Kong, N.T., Hong Kong
Tel: (852) 2609 8266 Fax : (852) 2603 5558
e-mail: {tlee1,pcching}@ee.cuhk.edu.hk

ABSTRACT

This paper describes a novel design of neural network based speech recognition system for isolated Cantonese syllables. Since Cantonese is a monosyllabic and tonal language, the recognition system consists of a tone recognizer and a base syllable recognizer. The tone recognizer adopts the architecture of multi-layer perceptron in which each output neuron represents a particular tone. The syllable recognizer contains a large number of independently trained recurrent networks, each representing a designated Cantonese syllable. Such a modular structure provides greater flexibility to expand the system vocabulary progressively by adding new syllable models. To demonstrate the effectiveness of the proposed method, a speaker-dependent recognition system has been built with the vocabulary growing from 40 syllables to 200 syllables. In the case of 200 syllables, a top-1 recognition accuracy of 81.8% has been attained and the top-3 accuracy is 95.2%.

1 INTRODUCTION

During the past decade, speech recognition technology has undergone significant progress and a number of practical applications have been successfully developed. Nevertheless, much remains to be done before machines can understand natural speech with human-like performance. Indeed, there exist multi-disciplinary problems and difficulties in speech recognition research. They are due to speech variabilities at different levels, some of which are well understood while the others are not [1]. On one hand, acoustic variabilities are caused by speaker variation, environmental condition, etc. These can be modeled effectively using statistical techniques, such as hidden Markov models (HMM) and artificial neural networks (NN). On the other hand, speech is language dependent and each language has its own properties. Such linguistic variabilities present an even more challenging problem since most existing methods have been developed mainly for spoken English and may not be appropriate for other languages, especially for those non-alphabetic ones like Chinese.

There are many different Chinese dialects. For historical reasons, Mandarin or Putonghua has been adopted as the "official" standard of spoken Chinese. In the area of speech recognition, Mandarin has also received most attention [2-4]. Among the other Chinese dialects, Cantonese has found its special importance of being used by tens of millions of people in Hong Kong and the rapidly developing Southern China economic zone. Moreover, Cantonese has a number of interesting linguistic features which make it very different from Western languages and even Mandarin. A good understanding of these features is important for the development of high performance speech recognition system.

Cantonese is a monosyllabic and tonal language. Each Chinese character is pronounced as a single syllable associated with a specific lexical tone. There are approximately 1,450 distinguishable tonal syllables being used in contemporary Cantonese. If the difference in tone is disregarded, the number of so-called base syllables is reduced to about 580. Each base syllable can be divided into an *Initial* and a *Final*. The *Initial* is an optional consonant while the *Final* can be a simple vowel, a diphthong, a vowel-nasal or vowel-stop combination. Cantonese has nine different tones which consists of six non-entering tones (labeled as $1-6$) and three entering tones ($7-9$). The entering tones occur exclusively in syllables with stop endings /p/, /t/ and /k/. Therefore they have short duration and abrupt energy drop at the syllable ending [5,6].

For speech recognition of a monosyllabic language, it is natural and reasonable to choose syllable as the basic recogniton unit. Automatic recognition of a Cantonese syllable requires parallel identification of the base syllable and the lexical tone. In our previous work, these two sub-problems have been studied extensively and a set of neural network based recognition techniques have been successfully developed [6-8]. In this paper, we shall describe the novel design of an integrated recognition system for isolated Cantonese syllables. Experimental results on speaker-dependent recognition of 200 commonly used Cantonese syllables will be reported to demonstrate the effectiveness of the proposed method.

2 THE SYSTEM DESIGN

As shown in Figure 1, the recognition system has two major components: a multi-layer perceptron based tone recognizer (TR) and a recurrent neural network based syllable recognizer (BSR). The outputs of TR and BSR are combined by an integrated algorithm to produce the ultimate recognition result.

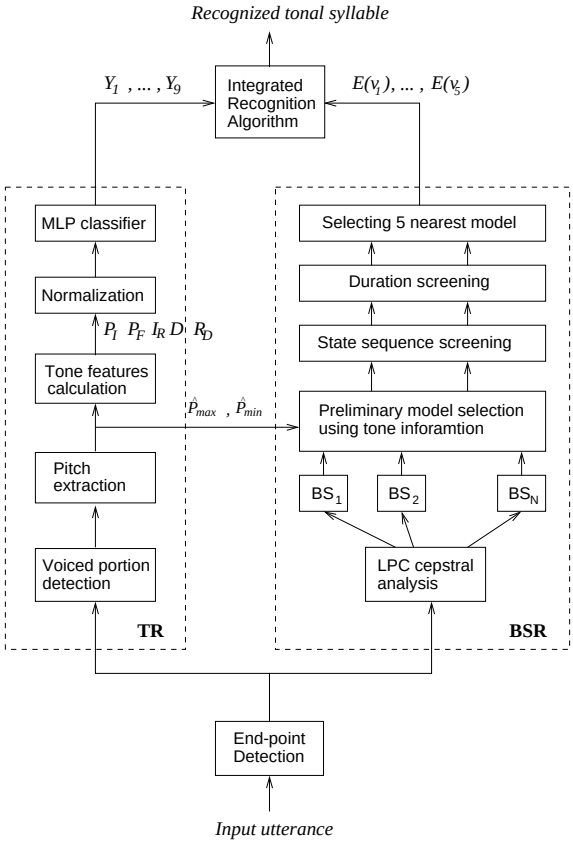

Figure 1: The proposed recognition system

Tone Recognizer (TR)

Tone is basically a feature of pitch movement over the entire syllable. For Cantonese in particular, tone identification relies on both pitch level and temporal pitch variation. Furthermore, tone duration and energy variation are also important to distinguish entering tones from non-entering tones.

In [6], we have proposed an effective method of using multi-layer perceptron (MLP) network for recognizing Cantonese lexical tones. The MLP contains three layers, namely the input layer, the hidden layer and the output layer. The input layer has 5 neurons, each accepting a supra-segmental feature parameter. These parameters include: *initial pitch* (P_I), *final pitch* (P_F), *pitch rising index* (I_R), *tone duration* (D) and *energy drop rate* (R_D). All of them are derived from the voice portion of the input syllable. A normalization procedure is employed to reduce undesirable variations of the feature parameters which may be caused by speaker difference and other unknown factors.

The output layer of MLP contains 9 neurons. Each of them represents a particular tone of Cantonese. Tone recognition is performed by the "winner-take-all" rule. That is, the output neuron with the highest activation level indicates the recognized tone.

In the recognition experiment with a tone-balanced speech corpus and 10 speakers (5 male and 5 female), an accuracy of 87.6% has been attained using the above method. This performance is found to be comparable with the results of human listening tests.

Base Syllable Recognizer (BSR)

While the tone recognizer utilizes multi-layer perceptron as a static pattern classifier, the base syllable recognizer is composed of dynamic recurrent neural networks (RNN). The recognizer input is a sequence of short-time spectral feature vectors. Each vector has 16 components which include energy, delta energy, LPC cepstral coefficients and delta cepstral coefficients.

For each base syllable, a fully connected RNN is trained independently to capture its static and dynamic features [7]. The occurrence of each phonetic component is represented by a particular output neuron in the RNN. To represent the temporal relation among these components, the output neurons are required to be activated one after another, following a specific sequential order. In addition, the activating periods of individual output neurons indicate duration of the corresponding phonetic components.

As shown in Figure 1, the BSR consists of a number of RNN syllable models. The recognition process is to identify the best matching syllable model for the input utterance. The reasons of adopting such a one-class-one-network modular structure are multi-fold. Firstly, we consider each Cantonese syllable as an unseparable and unique entity. A dedicated RNN can characterize its acoustic variation, both static and dynamic, in a more accurate and effective way. Secondly, training a large RNN is very time consuming and may have stability problem. By decomposing it into smaller subnetworks, the training efficiency is improved and the training process becomes more tractable so that unstable cases can be identified and handled. Thirdly, the modular structure offers greater flexibility to expand the recognition vocabulary. This can be done in a fairly straightforward manner by adding new syllable models and does not require re-configuration and re-training of the whole system.

Let the RNN models be denoted as BS_1, BS_2, $\cdots$, BS_N respectively. Given an input utterance, the best

matching base syllable is obtained via a multi-pass selection-by-elimination process:

Pass 1 – preliminary selection
Let $\hat{P}_{max}$ and $\hat{P}_{min}$ be the maximum and minimum normalized pitch which are obtained directly from the tone recognizer. For high and flat tones like tone 1 & 7, a fairly high pitch level is generally expected whilst tone 4, 6 & 9 have lower pitch. The two rising tones, i.e. tone 2 & 5, usually exhibit a large different between $\hat{P}_{max}$ and $\hat{P}_{min}$. Accordingly certain number of unlikely syllables can be eliminated by the following heuristic rules:

1) If $\hat{P}_{max} < 1.1$, the syllable could not be tone 1, 2 and 7;
2) If $\hat{P}_{min} > 1.5$, the syllable could not be tone 2, 4, 5, 6 and 9;
3) If $\hat{P}_{max} - \hat{P}_{min} > 0.6$, the syllable could not be tone 1, 3, 4, 6, 8 and 9;

Pass 2 – state sequence screening
This will eliminate syllable models in which the activation of output neurons fail to follow the desired sequential order

Pass 3 – duration screening
This will eliminate syllable models that violate the pre-determined segmental duration constraints. For example, in the entering tone syllable /sap/ ("ten"), the fricative /s/ should have a fairly long duration while the vowel /a/ must be very short. In practical applications, duration constraints are derived from training data.

Pass 4 – nearest model selection
For each eligible syllable model BS_n, an error function $E(n)$ is calculated as time average of the squared distance between the actual output and target output of the RNN. Then the model with the smallest $E(n)$ is selected as the recognition result.

To improve discrimination capability of the BSR, an MCE/GPD based discriminative training algorithm has been developed. Furthermore, the error function $E(n)$ is modified to equalize the effect of all phonetic constituents in the syllable, regardless of their duration difference [8,9].

Integrated Recognition Algorithm

To facilitate the design of an integrated recognition algorithm, the outputs of TR and BSR are re-formulated. For each input utterance, we consider the best 5 base syllable candidates from the BSR, which are denoted as BS_{v_1} to BS_{v_5} ($v_1, \cdots, v_5 \leq N$). The *a posteriori* probability of BS_n ($1 \leq n \leq N$) is estimated as

$$P_{BS}(n) = \begin{cases} \frac{[1/E(n)]^k}{\sum_{l=1}^{5}[1/E(v_l)]^k} & BS_n \text{ in the best 5} \\ 0.0 & \text{otherwise} \end{cases} \quad (1)$$

where $k \geq 1$ is a constant integer.

As for the TR, all of the nine tones are considered. Let Y_i denote the activation level of the ith output neuron ($0 \leq Y_i \leq 1$) in the MLP. The *a posteriori* probability of tone i ($1 \leq i \leq 9$) is calculated by

$$P_T(i) = \frac{Y_i}{\sum_{j=1}^{9} Y_j} \quad (2)$$

If BS_n and tone i together form a phonologically allowed Cantonese syllable, a probabilistic likelihood measure can be computed from

$$P_{TS}(n,i) = P_{BS}(n) \cdot P_T(i) \quad (3)$$

Then the tonal syllable with maximum value of $P_{TS}(n,i)$ is taken as the ultimate recognition result [5].

3 SIMULATION EXPERIMENTS

Currently we are working on a speech corpus of 200 Cantonese syllables which are commonly used to describe numbers, time, people, actions, etc. Counting all frequently used homonyms, the 200 syllables correspond to about 910 Chinese characters. Table 1 shows the proportions that these characters occupy in various text passages. In terms of phonetic coverage, it contains 19 *Initials* (out of 20), 44 *Finals* (out of 53) and all of the nine tones (Table 2). The total number of base syllables is equal to 166 (out of 580).

Three male speakers were asked to read all syllables in isolation. Each speaker contributed 12 sets of speech data, half of them (18 utterances for each syllable) are used for system training and the other half for performance evaluation.

As described in Section 2, the tone recognizer is a three-layer feed-forward network. In this particular application, 35 hidden neurons are used. The conventional error backpropagation algorithm is applied to train the MLP. The training is terminated if no further improvement can be attained for the training data.

A small-scale recognition system is constructed first for a sub-set of 40 syllables. For each base syllable, a fully connected RNN with 12 neurons is trained using an iterative re-segmentation algorithm [7]. This training algorithm does not require pre-segmentation of training data but attempts to estimate the optimal phonetic segmentation during the training process. After the independent training, discriminative training is carried out among the 40 RNN models until the recognition accuracy can not be further improved.

Newspaper text	Phrase dictionary	Governor's Address
39%	32%	48%

Table 1: Text coverage of the 200 Cantonese syllables

Initials	
L-C	IPA
b	p
d	t
g	k
gw	k^w
p	p^h
t	t^h
k	k^h
l	l
w	w
j	j
m	m
n	n
ng	ŋ
s	s/ʃ
f	f
h	h
dz	ts/tʃ
ts	ts^h/$tʃ^h$

Finals			
L-C	IPA	L-C	IPA
i	i	eng	ɛŋ
yu	y	eon	øn
u	u	oeng	œŋ
e	ɛ	ong	ɔŋ
o	ɔ	am	ɐm
aa	a	an	ɐn
ui	ui	aam	am
ei	ei	aan	an
eoy	øy	aang	aŋ
oi	ɔi	it	it
ai	ɐi	ik	ɪk
aai	ai	yut	yt
iu	iu	ut	ut
ou	ou	uk	ʊk
au	ɐu	eot	øt
aau	au	ok	ɔk
im	im	ap	ɐp
in	in	at	ɐt
ing	ɪŋ	ak	ɐk
yun	yn	aat	at
un	un	aak	ak
ung	ʊŋ	m	m

Table 2: A list of Cantonese phonemes covered in the speech corpus (L-C is a Cantonese romanization system adopted at Dept. of EE, CUHK)

By making use of the modular structure of BSR, we can expand the system vocabulary progressively to include 80, 120 and eventually cover the entire corpus. This is done by applying the following pragmatic training procedures:

(1) individual training of the new syllable models
(2) discriminative training among the new models
(3) discriminative training for the whole system

Step (1) & (2) cover basic training of the RNNs while step (3) represents the additional training due to the introduction of new syllable models. In our simulation experiments, it is observed that only a small number of training cycles are needed to complete step (3).

Table 3 shows the recognition performance of the proposed system. As the vocabulary being expanded, the top-1 recognition accuracy decreases significantly while the top-3 accuracy keeps at a fairly high level. In fact, a large proportion of the recognition errors are due to pairwise confusion of syllables. These syllables usually share the same or similar *Final* part. In some cases, even human listeners are unable to distinguish them reliably if no contextual information is provided. For example, /baat8/ ("eight") and /baak8/ ("hundred") are completely confused since the stop /t/ and /k/ are glottalized in Cantonese. If such completely confused syllables are grouped into the same class, the top-1 recognition rate is improved to 85.4%.

Vocab. size	Accuracy %		
	Top-1	Top-2	Top-3
40	96.9%	99.2%	99.9%
80	91.5%	98.3%	99.1%
120	88.1%	94.0%	97.3%
200	81.8%	89.8%	95.2%

Table 3: Recognition performance

Further investigation on expanding the vocabulary size and improving the recognition results are still undergoing. But this pioneer work on automatic recognition of Cantonese serves as a milestone for future development in this interesting area.

ACKNOWLEDGEMENT

This work is partly supported by research grants from the Croucher Foundation and the Hong Kong Research Grants Council.

REFERENCES

[1] L. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Inc., 1996.

[2] H.W. Hon *et al*, "Towards large vocabulary Mandarin Chinese speech recognition", *Proc. ICASSP-94*, Vol.1, pp.545 – 548.

[3] T.H. Ho *et al*, "Fast and accurate continuous speech recognition for Chinese language with very large vocabulary", *Proc. EUROSPEECH-95*, Vol.1, pp.211 – 214.

[4] B. Xu *et al*, "A general Chinese acoustic/phonetic decoder for syllable, word and continuous speech recognition", *Proc. ISSIPNN-94*, Vol.2, pp.706 – 709.

[5] Tan Lee, *Automatic Recognition of Isolated Cantonese Syllables Using Neural Networks*, PhD Thesis, The Chinese University of Hong Kong, May 1996.

[6] Tan Lee, P.C. Ching, L.W. Chan, B. Mak and Y.H. Cheng, "Tone recognition of isolated Cantonese syllables", *IEEE Trans. SAP*, Vol.3 No.3, pp.204 – 209.

[7] Tan Lee, P.C. Ching and L.W. Chan, "Recurrent neural networks for speech modeling and speech recognition", *Proc. ICASSP-95*, Vol.5, pp.3319 – 3322.

[8] Tan Lee, P.C. Ching and L.W. Chan, "An RNN based speech recognition system with discriminative training", *Proc. EUROSPEECH-95*, Vol.3, pp.1667 – 70.

[9] Tan Lee and P.C. Ching, "On improving discrimination capability of an RNN based recognizer", *Proc. ICSLP-96*, Vol.1, pp.526 – 529.