ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

INTEGRATING TONAL INFORMATION INTO MANDARIN NAME RECOGNITION WITH DIFFERENT STRATEGIES

Dong-sheng Luo, Xiang Xie, Jing-ming Kuang

Department of Electronic Engineering, Beijing Institute of Technology
{luodongsheng, xiexiang, jmkuang}@bit.edu.cn

ABSTRACT

Name recognition is a practical application of speech recognition technology. As Chinese is well known to be a tonal language, tonal information has important influence on this task. In this paper we integrate tonal information into a speaker-independent Mandarin name recognizer, and two combination strategies: feature combination and posterior combination are investigated firstly. The recognizer is evaluated on an extremely challenging Mandarin name corpus, which includes 100 tonally confusing pairs. Although a significant improvement in the recognition accuracy can be achieved with either strategy, the system has a poor flexibility. Based on the analysis of the experiment results we propose a two-step process to improve the system performance further. It is shown that a maximal improvement of 29.96% in word accuracy can be achieved. At the same time the system has a good flexibility with tonal information being integrated dynamically.

1. INTRODUCTION

As a practical application of ASR technology, name recognition has been used in many fields such as speech name dialer in a mobile phone. Different from some western languages, Chinese is a tonal language. Since there may be tonal confusions between Chinese names, tonal information plays an especially important role in this recognition task.

The pronunciation of a Chinese syllable depends on not only its phonetic pronunciation but also its tone. From the viewpoint of both speech production and perception, they have quite different characteristics. Therefore we can regard them as different sources of information in the speech signal and employ a multi-stream model to deal with them. In this way features extracted for speech recognition are divided into two separate feature streams: base feature stream and tonal feature stream. The base feature here is just the conventional cepstral feature, such as MFCC, with optional frame energy. To distinguish tonal difference between acoustic models, tonal feature is extracted. The commonly used tonal features include fundamental frequency, degree of voicing [1] and higher order cepstral coefficients [2].

The two feature streams can be combined at different levels: feature level or posterior probability level. Accordingly the two combination strategies are named feature combination and posterior combination. In this paper we compare the two combination strategies and propose a two-step process to improve the recognition performance further.

In the next section we describe our experimental setup and the database. Feature representation of the two streams and tonal feature extraction algorithm are described in section 3. In section 4 we introduce the two combination strategies and give the experimental results. Based on the analysis of the results, we further improve the performance with a two-step process in section 5.

2. EXPERIMENTAL SETUP

2.1. Acoustic Model Unit

Considering the availability of integrating tonal information, tonal semi-syllable is selected as the acoustic model unit in our experiments. We use preme/toneme [3] unit instead of traditional initial/tonal final unit. The concept of preme/toneme is based on the hypothesis that tonal information is concentrated in the pitch behavior of the main vowel of a final. Each preme is a combination of the initial and the glide of the final. Toneme is defined as the latter part of the tonal final, starting with the main vowel. Table1 gives some examples to show the difference between initial/tonal final and preme/toneme.

syllable	initial/tonal final	preme/toneme
lan4	l / an4	l / an4
lian2	l / ian2	li / an2
luan3	l / uan3	lu / an3

Table1: Different decomposition for Mandarin syllable: initial/tonal final and preme/toneme

Our preme/toneme unit includes 56 premes and 95 tonemes. Neutral tone is not taken into account here, since it seldom occurs in names. In addition there is an extra

models for silence, so there are totally 152 acoustic models in the experiments. Compared with initial/tonal final, preme/toneme unit has fewer models but is able to achieve similar recognition performance.

2.2. Database

In our experiments the "863" project LVCSR Mandarin corpus is selected as the training database. Here 86,453 utterances from 150 speakers (75 males and 75 females) are used to train a set of speaker-independent preme/toneme models.

According to the requirement of recognition task, a Mandarin name corpus was collected as the testing database. The vocabulary is composed of 50 pairs of short names (bi-syllabic names) and 50 pairs of long names (tri-syllabic names). Within a pair the two names are tonally confusing, which means that they have the same base syllables (syllable without tone) but different tones. Some examples are given in Table2. The database was contributed by 10 male and 10 female speakers, and each speaker was required to utter the 200 names twice in a quiet environment. The corpus constitutes an extremely challenging task, for there are tonally confusing pairs and the confusions can not be decreased by language model.

short name pair		long name pair	
顾萌	gu4 meng2	汪嘉伟	wang1 jia1 wei3
谷萌	gu3 meng2	王家卫	wang2 jia1 wei4

Table2: Examples of tonally confusing name pairs

2.3. Baseline System

A baseline speaker-independent Mandarin name recognizer has been implemented by HTK [4]. Each preme/toneme is modeled by a continuous HMM with 6 states and the left-to-right topology. The observation probability is represented by a Gaussian density with 4 mixtures. A single feature stream with 39 dimensions is extracted, and the feature vector consists of:

- MFCC feature (12MFCC, 12ΔMFCC, 12ΔΔMFCC)
- Energy feature (E, ΔE, ΔΔE)

All 20 speakers' 8000 utterances are used for testing, and short names and long names are tested together. The result is evaluated by two parameters: word accuracy with tone and word accuracy without tone. The former considers both the base syllables and tones in the utterances. However, the latter one neglects tones, in which the result is considered correct as long as the base syllable string is correctly recognized. The experimental result of the baseline system is shown in Table3. The figures indicate that the errors mainly come from the tonal confusion between names. The key for the system to increase its recognition accuracy is to improve the tonal distinguish ability.

	Word Accuracy(%)	
	with tone	without tone
Baseline	72.81	98.46

Table3: Recognition result of the baseline system

3. TONAL FEATURE EXTRACTION

3.1. Feature Representations

In our multi-stream system there are two feature streams: base feature stream and tonal feature stream. In the base feature stream the feature vector is just the same as that in the baseline system. As fundamental frequency (F_0) variation is the main discrimination feature of tones, here we use the normalized fundamental frequency (F_n) and its first-order and second-order derivative as tonal feature. Thus the feature vectors of the two streams consist of:

- Base feature stream: 12MFCC, 12ΔMFCC, 12ΔΔMFCC, E, ΔE, ΔΔE
- Tonal feature stream: F_n, ΔF_n, $\Delta\Delta F_n$

3.2. Fundamental Frequency Extraction Algorithm

In our experiments we use the pitch extraction algorithm introduced in [5]. It is a frequency-domain algorithm, and local peaks are firstly picked out from the power spectrum: $\{(A_k, n_k),\ k = 1,..., Npeaks\}$, where n_k is the location of local peak, A_k is the corresponding amplitude, and $Npeaks$ is the number of local peaks. Then F_0 candidates are selected among the local maxima of a piecewise constant utility function $U(F_0)$:

$$U(F_0) = \sum_k NA_k \times I(n_k / F_0) \tag{1}$$

where NA_k is the normalized A_k .

$$NA_k = A_k / \sum_{t=1}^{Npeaks} A_t \tag{2}$$

$$I(r) = \begin{cases} 1, & |r| \le D1 \\ 0.5, & D1 < |r| \le D2 \\ 0, & D2 < |r| < 0.5 \end{cases} \tag{3}$$

$$I(r+1) = I(r)$$

$$D1 = 65/512,\ D2 = 100/512$$

At last the correlation of each candidate is taken into account in addition to obtain the final F_0 estimation result.

3.3. Fundamental Frequency Post-process

The extracted F_0 cannot directly constitute feature vector, because it is meaningful only in those voiced frames of speech. In our experiments the F_0 in the silence parts is defined as the running average plus a random noise, and a liner interpolation is used to fill in those unvoiced frames. Previously a voiced/unvoiced detection is made in the fundamental frequency extractor as a by-product.

Another problem is that F_0 is speaker-dependent, and it is also a component of intonation. Here we use the long-term pitch normalization (LPN) [6] to eliminate the speaker and intonational effect.

4. BASIC COMBINATION STRATEGIES

After feature extraction when and how to combine the two feature streams is an important issue. In our experiments we compare two different basic combination strategies: feature combination (FC) and posterior combination (PC). The difference between them lies in the different stages of an ASR system where the two feature streams are combined. In feature combination tonal information is integrated after the feature extraction. However in posterior combination the two feature streams are combined in the pattern matching procedure.

4.1. Feature Combination

In this strategy feature vectors of each stream are concatenated to create a single, larger feature vector:

$$o^{(t)} = [o_B^{(t)}, o_T^{(t)}] \in R^D, \qquad D = D_B + D_T \tag{4}$$

where $o_B^{(t)}$ is the base feature vector of dimension D_B, $o_T^{(t)}$ is the tonal feature vector of dimension D_T. The training and testing are both based on the large feature vector.

4.2. Posterior Combination

In posterior combination separate acoustic models for each stream calculate the posterior probabilities, and the two probabilities are combined by some rules for later decoding process. Considering the acoustic meanings of the two feature streams, geometric weighted mean is chosen as the combination rule in our experiments. Thus the state emission probability of the multi-stream CDHMM are defined as follows:

$$b(o^{(t)}) = \prod_{s\in\{B,T\}} \left[\sum_{m=1}^{M_s} c_{jsm} N(o_s^{(t)}; \mu_{jsm}, \sum_{jsm}) \right]^{w_s} \tag{5}$$

with $0 \le w_s \le 1, \sum_{s\in\{B,T\}} w_s = 1$

where $s \in \{B, T\}$ represents base feature stream and tonal feature stream respectively, and w_s is the stream weight. The weights of the two streams are experiential according to the experimental results.

4.3. Experimental Result and Discussion

Feature combination (FC) and posterior combination (PC) have been tested in our experiments. The recognition results are shown in Table4.

From the results we can find that:

i. Significant improvement in the recognition performance is achieved as expected with either combination strategy.

	Word Accuracy (%)		Improvement (%)	
	with tone	without tone	with tone	without tone
Baseline	72.81	98.46	-	-
FC	91.05	98.45	25.05	-0.01
PC	93.94	97.83	29.01	-0.63

Table4: Recognition accuracy with different strategies

The testing database determines that the recognition accuracy mainly depends on the tonal distinguish ability. The conventional cepstral feature has only limited tonal information, and the integration of tonal feature solves the problem to a great extent. As a result the recognition accuracy increases significantly.

ii. There is a slight decrease in the base syllables accuracy. After analyzing the recognition results, we find that some utterances that can be correctly recognized in the baseline system, are recognized wrongly with tonal information integrated. There are two kinds of cases of these utterances: (a) The correct utterance and the recognized utterance have similar pitch contour. (b) In some frames F_0 is not correctly extracted. It is indicated that the integration of tonal feature maybe has some negative effects sometimes.

In both FC and PC tonal information is integrated all the time no matter whether it is needed by the situation. Besides the potential negative effect, it also increases the computation complexity of the recognizer. Therefore their flexibilities are not good.

5. IMPROVEMENT WITH A TWO-STEP PROCESS

5.1. System Structure

Based on the analysis above, we improve our name recognizer by proposing a two-step process. Now there are two recognizers in the system: a conventional recognizer and a tonally distinguishable recognizer. The two recognizer works successively, then the recognition task is separated into a two-step procedure as shown in Figure1. In step1 the base syllables are recognized using the base feature stream solely. Although the recognition result is also tonal syllable sequence, we think that only the base syllables are reliable. When tonal information is needed according to the recognition result in step1, tonal feature will be extracted and used in the tonally distinguishable recognizer. The necessity of step2 depends on the confusion degree between N-best candidates in step1, and here 3-best is considered.

The specific critical rules are as follows:

i. During the recognition set up, a set of tonally confusing pairs in the vocabulary can be established. If the 1st candidate in step1 falls into the set, which means its base syllables are shared by other name(s), the tonally confusing name(s) is regarded as confusing candidate(s) no matter it occurs in 3-best or not.

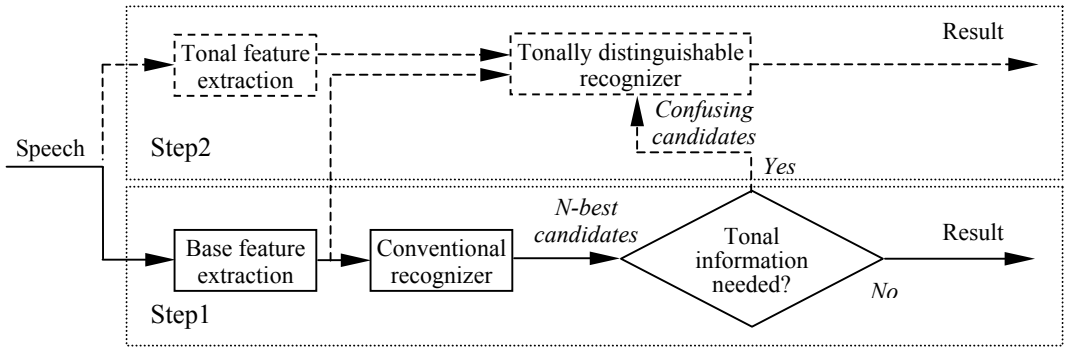

Figure1: System structure with a two-step process

ii. Among the 3-best, except the tonally confusing name(s), if a candidate's score difference between the 1st candidate is below an experiential threshold, it is also regarded as a confusing candidate.

As long as either criteria is validated Step2 will be implemented, and the 1st candidate together with the confusing candidate(s) compose the vocabulary in step2. In step2 tonal information has two functions. One is to distinguish tonally confusing pairs. The other is to correct base syllable errors in step1. The tonally distinguishable recognizer works as a secondary selector, and it runs based on both the two streams with posterior combination. Since the result of the baseline system has indicated that the base feature also contains some information for distinguishing tonal confusion, the integration of base feature is naturally helpful. At the same time, in the case of *ii* base feature is also need to distinguish confusing candidates. In our experiments stream weights of (0.2, 0.8) are used in step2 experientially.

5.2. Experimental Result and Discussion

The recognition result is shown in Table5. It is seen that the recognition accuracy has exceeded both FC and PC. This is attributed to the two-step procedure and thus the potential negative effect of tonal feature can be avoided. At the same time a higher accuracy of base syllables than the baseline is achieved as expected, which indicates that some base syllable errors in step1 has been corrected by tonal information. In this system tonal information is applied only when it is needed, and in other cases there is no extra computing consumption. The system with a two-step process shows a good flexibility.

	Word Accuracy (%)		Improvement (%)	
	with tone	without tone	with tone	Without tone
Baseline	72.81	98.46	-	-
Tow-step process	94.63	98.83	29.96	0.37

Table5: Recognition accuracy with two-step process

6. CONCLUSIONS

We have integrated tonal information into a speaker-independent Mandarin name recognizer to improve its performance. With feature combination and posterior combination, the recognition accuracy can be improved significantly, due to the tonal difference. However, their flexibilities are poor because tonal feature is used all the time. In the two-step process tonal information can be integrated dynamically according to the situation, and the negative influence and extra computing consumption are avoided in the needless cases. The system with the two-step process achieves the highest recognition accuracy and has a good flexibility. Therefore it seems to be a good solution to Mandarin name recognition task.

7. ACKNOWLEDGEMENT

The research was sponsored by the Beijing Institute of Technology Basic Research Foundation, and the cooperation project between Beijing Institute of Technology and Ericsson Group, Sweden.

8. REFERENCES

[1] D.L. Thomson et al, "Use of Periodicity and Jitter as Speech Recognition Features", in *Proc. ICASSP98*, Vol.1, pp21-24, Seattle, 1998.

[2] Xia Wang, Yuan Dong, Juha Häkkinen, Olli Viikki, "Noise Robust Chinese Speech Recognition Using Feature Vector Normalization and Higher-Order Cepstral Coefficients", *ICSP2000*, Beijing, China, 2000.

[3] C.J. Chen, R.A. Gopinath, M.D. Monkowski, M.A. Picheny and K. Shen, "New Methods in Continuous Mandarin Speech Recognition", in *Proc. Eurospeech 97*, Volume3, pages 1543-1546.

[4] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*, Cambridge, England, Microsoft, 2000, July.

[5] *ETSI ES 202 211 v1.1.1* (2003-11).

[6] H. Huang and F. Seide, "Pitch Tracking and Tone Features for Mandarin Speech Recognition", in *Proc. ICASSP2000*, vol. 3, pp. 1523-1526, 2000.