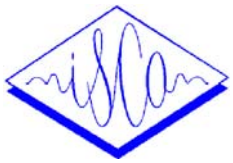
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Parallel Tone Score Association Method for Tone Language Speech Recognition

Gang Peng, William S-Y. Wang

Department of Electronic Engineering
City University of Hong Kong, Hong Kong
gpeng@ee.cityu.edu.hk (Peng) wsyw@ee.cityu.edu.hk (Wang)

10.21437/Interspeech.2004-167

Abstract

Tone is an essential component for word formation in all tone languages. Substantial work has been done on using tone information to improve speech recognition of tone languages. In this paper, a new method, called Parallel Tone Score Association (PTSA), for effectively and efficiently using tone in speech recognition is proposed. Experimental results show that the relative character error rates are reduced by as much as 20.94% for Cantonese, and 20.49% for Mandarin compared with the recognition results without tone information. This relative reduction in error rates compares favorably with results reported for other recognition experiments on tone languages.

1. Introduction

Tone languages form an important category of languages in the world. Tone is an essential component of tone languages, and is used to build words much as consonants and vowels do. For instance, in Mandarin, the syllable /ma/, when pronounced with a high level pitch pattern, means "mother"; when pronounced with a rising pattern, the meaning is "hemp"; when pronounced with a falling-rising pattern, the meaning is "horse"; when pronounced with a falling pattern , the meaning is "to scold" [4]. So speech recognition of tone languages depends not only on the articulatory composition but also on tone patterns.

During the last two decades, many approaches have been proposed for tone recognition. Hidden Markov Models (HMMs) and Neural Networks have been applied to recognize tones in tone languages, such as Mandarin and Cantonese. For isolated tone recognition, very high recognition accuracy has been obtained. However, automatic tone recognition in continuous speech is a difficult task, especially for a language such as Cantonese with a very complex tone system.

How to integrate tone recognition into speech recognition is an important problem, which is the focus of this paper. One approach recognizes tonal syllables directly using the HMMs. In this approach, the tone related features are included as augmented components of the acoustic feature vector. However, this approach requires a large amount of training data, because syllable finals with different tones cannot be shared in training. Moreover, this approach may introduce extra correlations among different components of the feature vector, which may impede the using of diagonal covariance matrix for acoustic modeling to reduce the computational cost. For Mandarin, positive results have been reported in [1, 2]. However, when applied to Cantonese speech recognition, adding tone information like this actually lowered the recognition score (reported as an additional test in [3] for their comparison). The second approach recognizes tones separately from articulatory recognition. This approach can be further divided into two sub-approaches. The first sub-approach separates the whole recognition task into two steps. In the first step, the base syllables (without considering tonal difference) are recognized. Then in the second step, tones of the recognized base syllables are identified with F_0 values and other tone related features. The main disadvantage of this sub-approach is that errors in the base syllable recognition cannot be recovered during the tone recognition, which results in locally optimal searching. The second sub-approach is to incorporate tone contribution in parallel to lattice generation. Our approach belongs to this category.

In this paper, an innovative approach, called Parallel Tone Score Association (PTSA), will be presented to incorporate tone contribution in parallel to lattice generation. In the next section, the tone systems of Mandarin and Cantonese will be briefly described. The method of PTSA will be introduced in section 3. Experimental results will be presented in section 4. Finally, conclusions will be drawn in section 5.

2. Tone systems of Mandarin and Cantonese

Cantonese has a rich inventory of tones. Traditionally, it is said there are nine lexical tones in Cantonese. Tones 1-6 occur are long tones because they occur on syllables without stop endings; these are called unchecked syllables. Tones 7, 8 and 9 are short tones because they alone occur on syllables with stop endings; these are called

checked syllables. Since their F_0 values correspond to the long tones 1, 3 and 6 respectively, they are numbered according to their long tone counterparts. This is done in many transcription schemes, including that of the Linguistic Society of Hong Kong, where only six distinctive tones are labeled. In Mandarin, there are only four lexical tones and one neutral tone whose F_0 contour completely depends on its immediately preceding tone, which is a highly context-dependent tone.

3. Parallel Tone Score Association

In this section, we will first summarize our tone recognition schemes. Then the method of PTSA will be introduced.

3.1. tone recognition

Since the tone recognition has been discussed in detail in [5], we simply discuss the tone features we use here. The tone of a syllable is mainly determined by its F_0 contour. The duration and energy are also related to the tone. For tone recognition in continuous speech, including context information from the neighboring tones improves the recognition accuracy of tones [6]. For a given syllable, the tone-related feature vector, consists of the following 20 features in our tone recognition schemes.

(1) Duration of the F_0 contour of the target syllable; F_0 values at both the 1/3 and 2/3 time points of each of the three uniformly divided linearly-fitted F_0 sub-contours; the means of the three corresponding log-energy sub-contours.

(2) The same three features (i.e., two F_0 values, mean of the log-energy) of the last sub-segment of the preceding F_0 contour and the corresponding log-energy sub-contour, and the first sub-segment of the following F_0 contour and the corresponding log-energy sub-contour.

(3) Log-energy and duration of unvoiced/silent segments both before and after the target syllable.

The ten features in (1) are all extracted from the target syllable. The six features in (2) are used to consider the tone coarticulation effect from the neighboring tones, (Because during online processing, the timing of the following syllable is unknown, we take the first following ten F_0 values as the first sub-segment of the following F_0 contour. To keep symmetry, we also take the last preceding ten F_0 values as the last sub-segment of the preceding F_0 contour.) while the 4 features in (3) are used to implicitly represent the tightness of the coupling between the target tone and its neighboring tones. This feature selection scheme is similar to [6]. In our schemes, for Cantonese, their $slope$ feature of each linearly-fitted F_0 sub-contour is discarded, and their F_0 mean replaced by two F_0 values for each linearly-fitted F_0 sub-contour; while for Mandarin, the two F_0 values at both the 1/3 and 2/3 time points of each F_0 sub-contour are replaced by the corresponding $slope$ feature and the mean value.

3.2. PTSA method

The basic idea is to add tonal contribution in parallel to syllable lattice generation. When a syllable in one path reaches its end state, it may be added to the syllable lattice (and may be omitted by beam pruning). If it is added to the syllable lattice, then the tone features for this target syllable will be extracted on-line (If the recognized syllable is silence, then there is no need to expand it.). Those tone features will be immediately fed to the tone classifiers proposed in [5] (In [5], three tone recognition schemes have been proposed. The basic scheme is used here which has the lowest tone recognition accuracy among the three schemes. Thus the effectiveness of PTSA can be more prominent.). Meanwhile, the target syllable will be expanded to all possible tonal syllables (Illegitimate tonal syllables will be omitted.). The diagram of this expansion is shown in Fig. 1. How to distribute the total tonal contribution of the target base syllable to each tonal syllable is the crucial point of the tonal syllable lattice generation. Because only voiced frames have meaningful F_0 values, and F_0 is by far the most important manifestation of tones, we define the total tonal contribution over the voiced frames as

$$S_{Tone} = C \times L, \tag{1}$$

where C is a language related constant, and L is the number of voiced frames of the target syllable (30 was selected empirically for both target languages here.). If S_{Tone} is equally distributed to each tonal syllable, then no tonal syllables will be preferred; consequently, an equal amount of tonal score will be added to each path of the lattice. But in PTSA, each tone in a tone language is assigned a recognition score by the tone classifier based on the tone-related feature vector, and then S_{Tone} will be distributed to each tonal syllable proportionally according to its tonal recognition score as

$$P(T_iS) = P(S) + \frac{P(T_i)}{\sum_{j=1}^{N} P(T_j)} \times S_{Tone}, \tag{2}$$

where the $P(T_iS)$ is the probability of the tonal syllable with Tone i; $P(S)$ is the probability of the base syllable S (articulatory recognition score); $P(T_i)$ is the probability of Tone i (tonal recognition score); and N is the number of tones of the target language. In this way, the tonal score is associated with the articulatory score in parallel.

4. Experimental results

In this section, the PTSA method will be evaluated for Cantonese and Mandarin Large Vocabulary Continuous

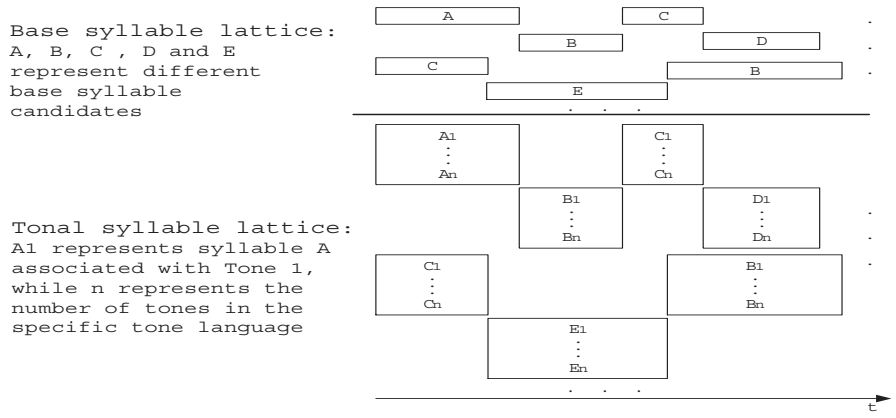


Figure 1: *Base syllable lattice expansion with tone information.*

Speech Recognition (LVCSR) tasks, respectively. Cantonese has one of the richest tone systems in the world, while Mandarin is a typical tone language [4]. If PTSA can deal well with the integration of tone recognition with articulatory recognition in these tone languages, then this approach should be useful for other tone languages.

4.1. Database

The Cantonese database we used is CUSENT database [7]. In this database, 5,100 training and another exclusive 600 test sentences were separately selected from five local newspapers of Hong Kong. The training sentences were evenly divided into 17 groups, each containing 300 unique sentences. Each group of sentences were read by four speakers (2F, 2M). Thus, a total of 20, 400 (300x4x17) training utterances were obtained from 68 speakers. The 600 test sentences were divided into 6 groups. Each group was read by one male and one female speaker (not drawn from the population of the training speakers). The total number of test utterances is 1,200.

For Mandarin database, the experiments are performed on the later developed part of the database from Chinese Project 863. 1,560 sentences were selected from "The People's Daily". They are divided into three groups: group A with 521 sentences, group B with 519 sentences, and group C with 520 sentences. Group A were read by 27 (13F, 14M) speakers; group B by 28(14F, 14M) speakers; and group C by 27 (14F, 13M) speakers. The speakers for the above three groups are exclusive. The utterances from randomly selected 6 (3F, 3M) speakers are used for testing. Table 1 summarized the Cantonese and Mandarin database. Please note some corrupted utterances are excluded from this table.

4.2. Baseline system

The acoustic models consist of context-dependent initial-final models (triphone models), in which each initial model has 3 emitting states, while a final model has either three or five emitting states, depending on its articulatory composition. Each emitting state consists of 8 Gaussian mixtures. The acoustic feature vector has a total of 39 components, including 12 Mel-Frequency Cepstral Coefficients (MFCCs), energy, their first-order derivatives and second-order derivatives. The Cantonese and Mandarin HMMs were trained with the above Cantonese and Mandarin database respectively. Decision tree-based clustering methods were used to facilitate sharing of model parameters. Base syllable recognition accuracies of 79.08% and 86.87% were obtained for the test sets for Cantonese and Mandarin respectively.

Table 1: *Training and test data for Cantonese and Mandarin.*

Properties	Cantonese		Mandarin	
	Training Data	Test Data	Training Data	Test Data
#speakers	68 34F,34M	12 6F,6M	76 38F,38M	6 3F,3M
#syllables	215,604	11,677	510,791	40,334
#utterances	20,378	1,198	39,519	3,120

The Cantonese language model, character-based trigrams, has been built with 3,927 character entries, which cover 99.99% of the training Cantonese text corpus. The training text contains about 150 million Chinese characters from Local Hong Kong newspapers. Using the test data of CUSENT and the above language model, a character accuracy of 80.90% was obtained without using tone information. The Mandarin character-based trigrams have been built with 3,999 character entries, which cover 99.9% of the training Mandarin text corpus. The training text contains about 153 million Chinese characters obtained from Tsinghua University in Beijing. Using the test data of Mandarin database and the above language model, a character accuracy of 89.41% was obtained without using tone information.

Table 2: *Cantonese system performance comparison with another current system*

	Base syllable accuracy	Character accuracy			
		Without tone	With tone	Improvement	Relative improvement
Lee et al's system [3]	75.69%	75.43%	76.61%	1.18%	4.80%
PTSA system	79.08%	80.90%	84.90%	4.00%	20.94%

Table 3: *Mandarin system performance comparison with other systems*

Systems	Improvement
Cao et al's system [8]	15.5%
Wang's system (YINHE domain) [9]	15.9%
PTSA system	20.49%

4.3. Experimental results

As shown in Table 2, with the same database, i.e., CUSENT, the PTSA system has achieved a relative improvement of as much as 20.94%, which significantly outperforms the system reported in [3].

As for Mandarin, the character accuracy has been improved from 89.41% to 91.58% after using tone information with PTSA method. Extensive work has been done for integrating tone information into Mandarin speech recognition. Table 3 compares performance gain in terms of relative improvement. Among the three systems, which all recognize tones separately from articulatory recognition, PTSA system outperforms all other systems.

5. Conclusion

Significant improvement for Cantonese and Mandarin speaker-independent LVCSR tasks have been achieved by incorporating tone information via the PTSA method. Compared with other methods of integrating tone into speech recognition, PTSA is more powerful in making use of tone information. There are mainly three aspects of the PTSA method that we would like to highlight. First, PTSA does not make any hard decision for tone recognition. Thus the error in tone recognition will not introduce immediate error in speech recognition. This is consistent with the probabilistic framework of HMMs. Second, the tone decoding is parallel to articulatory decoding by HMMs, which may avoid deletion of the right candidates by articulatory recognition alone. Third, PTSA is a general framework for incorporating tone information into speech recognition. It can be used to integrate tone information for other tone languages such as Thai.

Some of the mentioned aspects are shared by other related methods. However, these aspects have no significance unless good results can be achieved. The most important thing is that PTSA produced the highest performance improvement among the reported tone incorporation approaches. In the study reported here, the features selected for tone recognition have not yet been optimized. Exploiting the full potential of tone information to improve the tone recognition performance will doubtless lead to further improvement of the innovative approach presented here.

6. References

[1] Wong, P. F. and Siu, M. H., "Integration of tone related feature for Chinese speech recognition", in Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces, 64-68, 2002.

[2] Huang, H. C.-H. and Seide, F., "Pitch tracking and tone features for Mandarin speech recognition", in Proceedings of the ICASSP, 1523-1526, 2000.

[3] Lee, T., Lau, W., Wong, Y. W. and Ching, P. C., "Using tone information in Cantonese continuous speech recognition", ACM Transactions on Asian Language Information Processing, 1(1): 83-102, 2002.

[4] Wang, W. S-Y., "The Chinese language", Scientific American, 288:50-60, 1973.

[5] Peng, G. and Wang, W. S-Y., "Tone recognition of continuous Cantonese speech based on support vector machines", Speech Communication, submitted for publication.

[6] Chen, S. H. and Wang, Y. R., "Tone recognition of continuous Mandarin speech based on neural networks", IEEE Transactions on Speech and Audio Processing, 3(2):146-150, 1995.

[7] Lee, T., Lo, W. K., Ching, P. C. and Meng, H., "Spoken language resources for Cantonese speech processing", Speech Communication, 36(3-4):327-342, 2002.

[8] Cao, Y., Deng, Y. G., Huang, T. Y. and Xu, B., "Decision tree based Mandarin tone model and its application to speech recognition", in Proceedings of the ICASSP, 1759-1762, 2000.

[9] Wang, C., "Prosody modeling for improved speech recognition and understanding", Ph.D. dissertation, MIT, 2001.