RECOGNIZE TONE LANGUAGES USING PITCH INFORMATION ON THE MAIN VOWEL OF EACH SYLLABLE

C. Julian CHEN

IBM Thomas J. Watson Research Center, Yorktown Heights, 10598 NY, USA

juchen@us.ibm.com

Haiping LI, Liqin SHEN, GuoKang FU

IBM China Research Lab, Shangdi, Beijing 100085, China
lihp, shenlq, fugk@cn.ibm.com

ABSTRACT

An innovative method for speech recognition of tone languages is reported. By definition, the tone of a syllable is determined by the pitch contour of the entire syllable. We propose that the pitch information on the main vowel of a syllable is sufficient to determine the tone of that syllable. Therefore, to recognize tone languages, only main vowels are needed to associate with tones. The number of basic phonetic units required to recognize tone languages is greatly reduced. We then report experimental results on Cantonese and Mandarin. In both cases, using the main vowel method, while the number of phonemes and the quantity of training data are substantially reduced, the decoding accuracy is improved over other methods. Possible applications of the new method to other tone languages, including Thai, Vietnamese, Japanese, Swedish, and Norwegian are discussed.

1. INTRODUCTION

Tone language is an important category of languages, where pitch is a distinguishing mark of morphemes. Syllables or words having the same sequence of consonants and vowels but different pitch contours are often different lexical entries. Examples of tone languages include Chinese languages (such as Mandarin, Cantonese, Taiwanese), South Asia languages (such as Thai, Vietnamese), Japanese, Swedish, and Norwegian. Speech Recognition of tone languages depends not only on the phonetic composition but also on the lexical tone pattern.

In 1980s and early 1990s, a popular method for recognizing tone languages is the two-step method [1]. First, to recognize the base syllable by its consonants and vowels. Second, to recognize the tone of the syllable by classifying the pitch contour of that syllable using discriminative rules. The recognition of toned syllables is a combination of the recognition of base syllables and the recognition of tones. The above method works well in isolated-syllable speech recognition. But, it shows difficulties in handling continuous speech.

In mid 1990s, the one-step method for recognizing continuous tone languages is proposed [2]. In its early implementations, it uses a demisyllable approach. According to that approach, each syllable is decomposed into two demisyllables. The first demisyllable does not contain tone information. The second demisyllable, called toneme, carries the tone information of the whole syllable. In such a system, the second demisyllable with different tones are defined as different phonemes. It leads to a successful continuous Mandarin speech recognition system.

However, when generalizing it to tone languages with a large number of finals and tones than Mandarin, such as Cantonese, more than 300 phonemes are required. The larger the phoneme set, the more training data is required. Decoding is slowed down because of the large searching space.

Based on our observations of the tone patterns in Chinese languages, we found that it is plausible that the pitch information on the main vowel alone is sufficient to determine the tone of the whole syllable. Because the number of different main vowels is small (9 in Mandarin and 10 in Cantonese), if it is true, then the total number of phonemes required to recognize tone languages can be substantially reduced. However, the assumption that the pitch information on the main vowel alone is sufficient to determine the tone of the whole syllable has never been verified. Practically, that assumption can be verified by speech recognition experiments. If the reduction of the number of phonemes does not cause deterioration of decoding accuracy, that assumption is verified. Our results on both Cantonese and Mandarin show that its accuracy is even better than the demisyllable scheme, instead of being deteriorated [3].

The new method is universal. In other words, it can be applied to all tone languages. We will discuss possible applications to other tone languages.

2. THE ONE-STEP METHOD

In contrast to the two-step method, the one-step method recognizes vowels and consonants and tones in a single step, as shown in Fig. 1. As shown, pitch is treated as a component of the acoustic feature vector, the same way as cepstral coefficients and energy. Basic acoustic units with different tones are treated as different phonemes. Since tone is a property of syllable, a logical choice of the basic units is to use the syllables with tones. In Mandarin, there are about 1400 such units. It is too large to handle.

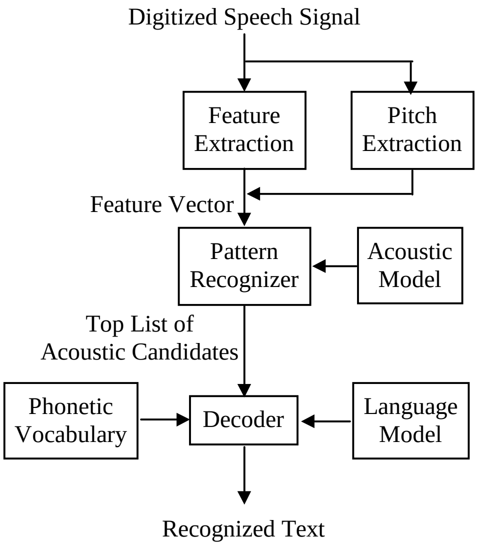


Figure 1. Speech Recognition System for Tone Languages

3. THE DEMI-SYLLABLE APPROACH

It is a standard practice to decompose each Chinese syllable to a consonant initial (Shengmu) and a final (Yunmu). The consonant initials do not contain tone information. Finals contain tone information. For example, in Mandarin, we have the following decomposition:

Syllable	Initial	Final
Lan3	L	AN3
Lian3	L	IAN3
Lang3	L	ANG3
Liang3	L	IANG3

In the Pinyin system for Mandarin, there are 21 initials and 39 finals. Including the neutral tone, each final could have five toned versions. By treating each initial and each toned final as phonemes, the total number of phonemes is

$$\boldsymbol{N} = \boldsymbol{C} + \boldsymbol{T} * \boldsymbol{V} , \qquad [1]$$

where $\boldsymbol{N}$ is the total number of phonemes, $\boldsymbol{C}$ is the number of consonant initials, $\boldsymbol{T}$ is the number of tones, and $\boldsymbol{V}$ is the number of vowel endings. For Mandarin, the total number of phonemes required is

20+5*39=215.

As shown in Reference [2], by grouping the glide with the consonant initial to become a preme, the total number of phonemes is reduced to

53+5*21=158.

Here, 53 is the number of premes, that is, combinations of initial consonants and glides; 5 is the number of tones; and 21 is the number of the second demisyllable without the glide, as shown by the following table:

Syllable	Preme	Toneme
Lan3	L	AN3
Lian3	LI	AN3
Lang3	L	ANG3
Liang3	LI	ANG3

That total number of phonemes, 158, is much smaller than the 215 phonemes using the initial-final scheme. However, for Cantonese, even with such simplifications, the total number of phonemes exceeds 300.

4. THE MAIN VOWEL APPROACH

With the assumption that the pitch information on the main vowel is sufficient to determine the tone of the whole syllable, the number of phonemes can be dramatically reduced. The above table now becomes:

Syllable	Phone 1	Phone 2	Phone 3	Phone 4
Lan3	L	A3	N	
Lian3	L	Y	A3	N
Lang3	L	A3	NG	
Liang3	L	Y	A3	NG

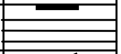
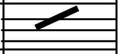
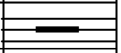
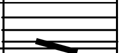
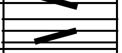
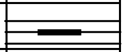
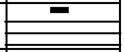
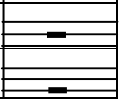
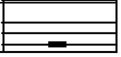
Because the number of different main vowels in Mandarin is only 9, the total number of phonemes is greatly reduced. The smaller phone set is quite advantageous. With the same amount of training data, a smaller phoneme set can be trained more thoroughly. When decompose syllables into phonemes, there are always some phonemes that only exist in very few syllables. Such phoneme is called a rare phone. The acoustic models for the rare phonemes are usually not so robust as the phonemes with high frequencies, because it's difficult to collect sufficient training data. If a much smaller phoneme set can be used, then the rare phonemes are much fewer or even do not exist. Therefore, it is much easier to design training scripts, to do training and decoding. In addition, phonemes from the new method are closer to the ones of Western languages. This makes it much easier to build multilingual speech recognition systems.

A question: Is the assumption that the main vowel contains sufficient tone information valid? Especially, for example, the A's in the above table are slightly different. Is the context-dependence tree building mechanism sufficient to handle it? To answer the question, we did several extensive experiments with Cantonese and Mandarin. We have positively verified the above assumption.

5. CANTONESE

Cantonese is spoken in the Province of Guangdong and Hong Kong. Its phonology consists of 20 initials and 59 finals. See Table 1.

Table 1. Tones in Cantonese

Groups	Tone Num.	Tone Name	Five Level System	Pitch Contour
Non-Entering Tones	1	High Level	55 or 53	
	2	High Rising	35	
	3	Mid Level	33	
	4	Low Falling	21	
	5	Low Rising	23	
	6	Low Level	22	
Entering Tones	7	High Entering	55	
	8	Mid Entering	3 or 33	
	9	Low Entering	2	

The most common numbering of Cantonese tones is 1 through 9. Tones 1-6 are for open syllables (so-called non-entering tones). Tones 7-9 are for closed syllables with plosive syllable endings p, t, k (so-called entering tones). However, the pitch values of tones 7, 8, and 9 are similar to those of tone 1, 3 and 6, respectively. The only difference is the length of the vowels. According to the Romanization Scheme of the Linguistic Society of Hong Kong [4], the nine tones can be combined into six. In the demisyllable approach, even using such a simplification, according to Eq. [1], the number of phonemes is

20+6*59 = 374.

In the main-vowel approach, only the following ten main vowels have tone distinction:

A AA E I O EU U YU M NG.

Adding the initial consonants

B CH D F G GW H J K KW L M N NG P S T W Y,

and the codas

P T K N NG M W Y,

the total number of phonemes, according to Eq. [1], is

28+6*10=88.

Two extensive speech recognition experiments have been conducted to verity the validity of the main-vowel method. Two sets of training data and test data are collected:

1). The small training set, collected from Guangdong province, contains 18,857 sentences. The test set has 4 male speakers and 4 female speakers, 100 utterances for each speaker.

2). The large training set, collected from Hong Kong, contains 29,964 sentences. The test set also has 4 male speakers and 4 female speakers, 100 utterances for each speaker.

Sharing the same language model, we compared the performance of two different acoustic models built on two phone sets. One is generated by the demisyllable method using the tone information of the finals. Another one is generated by the method of using the pitch information of the main vowel of the syllable. Since word-based language model is used, the recognition accuracy of the output text is calculated on the character level instead of phonetic syllables. The recognition error rate is listed in Table 2.

Table 2. Cantonese Test Results

Method of Tone Recognition	Training/test data from Guangdong	Training/test data from Hong Kong
Demisyllable	15.99	12.92
Main vowel	14.89	12.71

As shown, for both sets, the accuracy using the main-vowel method is better than the demisyllable method, despite that the size of phoneme set is significantly reduced. The improvement for the small training set is much greater than that of the small training set. This is expected that when the training set is small, the rare phone problem is more severe.

6. MANDARIN

As shown in Section 3, by using the textbook initial-final scheme, the number of phonemes required is 215; while an improved demesyllable method reduces the number to 158. Using the main-vowel method, the total number of phonemes can be further reduced by more than a factor of 2. In fact, in Mandarin, there are 9 main vowels:

A E EH ER I IH O U YU,

21 initial consonants,

B P M F D T N L Z C S ZH CH SH R J Q X G K H,

three glides,

W Y YU,

and three codas

W Y N NG.

The total number of phonemes, according to Equation [1], is

38+9*5=73.

Experiments are also performed on Mandarin to compare the performance of two systems built on different schemes. The training data for the acoustic model based on the demisyllable method includes more than one hundred thousands of continuous sentences. That for the min-vowel method includes about 40,000 sentences. Though the latter has much less training data, its accuracy is even better than that based on a larger training set using the demisyllable method, because the number of phonemes is much less than the demisyllable-based system.

For both cases, there are 20 test speakers in the test set, including 11 females and 9 males, and each speaker has 100 sentences. The result is shown in table 3.

Table 3. Mandarin Test Results

Method	Error Rate (%)
Demisyllable	13.65
Main vowel	12.61
Without pitch	16.79

For comparison, the result of a system without using pitch is also included. As shown, without pitch, the error rate is much larger than that of either of the two methods using pitch information.

7. OTHER TONE LANGUAGES

The method described here is universal, which can be applied to any tone language. As examples, we discuss applications to Thai, Vietnamese, Japanese, Swedish, and Norwegian.

7.1 Thai

Standard Thai has five phonemic tones, namely, mid, low, falling, high, and rising. It has 18 single vowels, 3 diphthongs, 21 single initial consonants, 12 initial consonant clusters, and 8 codas. The codas are identical to those in Cantonese (P T K N M NG W Y). There are several possible ways to implement the main-vowel method. By treating all vowels including diphthongs are main vowels, and all other phonemes as independent, the total number of phonemes is 41+5*21=226. However, the diphthongs can be decomposed into a strong vowel and a weak vowel, and the initial consonant clusters can be decomposed into to a consonant and a glide (semivowel). There are overlaps among consonants, codas, and semivowels. The minimal number of phonemes required is 21+5*18=111. Schemes in between can also be established. The best choice should be determined through large-scale experiments.

7.2 Vietnamese

The general syllable structure of Vietnamese is similar to that of Cantonese and Thai. It has 25 initial consonants, 11 single vowels, and eight consonant endings. However, a large number of vowel clusters can be formed. Thus, the main vowel method is particularly useful. Vietnamese has six tones. Using the main vowel method, the total number of phonemes is 33+6*11=99.

7.3 Japanese

The role of pitch in Japanese is quite different from other Asia languages such as Chinese, Thai, and Vietnamese. In Japanese, the pitch of each syllable could be either high or low. The pattern of pitch in a word is phonemic. For example, by denoting a syllable with high pitch with bold-faced letters,

hashi	is	chopstick
ha**shi**	is	bridge
hashi	is	end

Using the main-vowel method, to recognize tones in Japanese, only the five vowels (a, i, u, e, o) need to have high-pitch and low-pitch versions.

7.4 Swedish and Norwegian

The role of pitch in Swedish and Norwegian is yet different from all the above. For many multi-syllable words in those two languages, there are two pitch contours, or two tone patterns, which are called Tone 1 and Tone 2, respectively. Words with the same consonants and vowels but different pitch contours often have different meanings. For example, in Swedish,

"anden" pronounced as "an'den" is the duck

"anden" pronounced as "an`den" is the spirit

In Norwegian,

"badet" pronounced as "ba`de" is the bathroom

"bade" pronounced as "ba'de" is bathe

Thus, by assigning a rising tone and a falling tone to each vowel, the tones in those two languages can be recognized.

8. CONCLUSIONS

By extensive speech-recognition experiments on Cantonese and Mandarin, we show that the pitch information on the main vowel of a syllable is sufficient to determine the tone of that syllable. Based on that fact, simple and accurate speech recognition systems for tone languages can be developed which requires little training data. This method can be applied to all tone languages.

The authors wish to thank James Yeh for his constant support of this research project.

9. REFERENCES

[1] H. M. Wang, J. L. Shen, Y. J. Yang, C. Y. Tseng, and S. L. Lee, "Complete Chinese dictation system research and development", Proceedings ICASSP-94, Vol. 1, pp. 59-61.

[2] C. J. Chen, R. A. Gopinath, M. D. Monkowski, M. A. Picheny, and K. Shen, "New Methods in Continuous Mandarin Speech Recognition", 5th European Conference on Speech Communication and Technology, Vol. 3, pp. 1543 - 1546, 1997.

[3] Haiping Li, Liqin Shen, Guokang Fu, and C. Julian Chen, "A Promising Syllable Decomposition Method for Tonal Language's Speech Recognition", Proceedings of ISCSLP 2000, October 2000, Beijing.

[4] Guo Fang, Xingde Li, Sun Lin, Jingguang Lu, Zhesheng Tong, Qunxian Zhang, "Yue Yu Pin Yin Zi Biao", the Linguistic Society of Hong Kong, 1st ed., Jun. 1997.