SEGMENTAL TONAL MODELING FOR PHONE SET DESIGN IN MANDARIN LVCSR

Chao Huang, Yu shi, Jianlai Zhou, Min Chu, Terry Wang and Eric Chang

Microsoft Research Asia
5F, Sigma Center, No. 49, Zhichun Road, Beijing 100080, P.R. China
{chaoh, yushi, jlzhou, minchu, terryw, echang}@microsoft.com

ABSTRACT

Modeling units play a very important role in state-of-art speech recognition systems. The design and selection of them will directly impact the performance of final speech recognition engine. As a tonal language, Mandarin's modeling units are more special for the tonal processing. In this paper, after fully investigating several dominant modeling strategies, we propose a new phone set design strategy for Mandarin, called segmental tonal modeling. Instead of modeling tone types directly, we realized them implicitly and jointly by two segments, which both carry tonal information. Both HTK and SAPI based experiments confirmed that such method is very efficient. In addition to improving the accuracy by 9~23%, it greatly reduces the decoding time by 30~45%. Given the similar decoding speed, new phone set configuration can reduce the error rate by relatively 35%.

1. INTRODUCTION

Selecting the most suitable units to represent salient acoustic and phonetic information for a language is an important issue in designing a workable speech recognizer. In other word, modeling units play a very important role in state-of-art speech recognition. Usually, there are several criteria for choosing the appropriate modeling units.

- Accurate and representative: in different context;
- Trainable: the parameters of units can be estimated reliably with enough data;
- Generalizable: new words can be easily derived from the predefined unit inventory.

Correspondingly, there are several layers of units to be considered: phones, syllables and words. Their performances in term of above criteria are very different [1]. Word-based units should be a good choice for domain specific, such as recognizer designed for digits. However, for LVCSR, phone-based units (e.g. about 50 phones for English) are reasonable since they are more trainable and generalizable. Context-dependent phones, like tri-phone, are proved to be accurate together with state-sharing technology. Furthermore, there is another trend in state-of-art systems. Different levels of modeling units are used together in a system. For example, whole-word modeling is successfully integrated into phone based LVCSR system and states between them can also be shared.

Compared with most western languages, there are several distinctive characteristics for Chinese Mandarin:

1. The number of words is unlimited while number of characters and syllables are fixed. Specifically, one Chinese character corresponds to one syllable. Totally, there are about 420 base syllables and more than 1200 tonal ones. Entire tonal-syllable inventory construct the whole pronunciations of Mandarin.
2. Chinese is a tonal language. For each syllable, there are usually 5 tone types from tone 1 to tone 5, like *{/ma1/ /ma2/ /ma3/ /ma4/ /ma5}/*. Although the phones are the same, the real acoustic realizations are different because of the tone types as shown in Figure 1. It will become one of the main issues when designing the phone set.

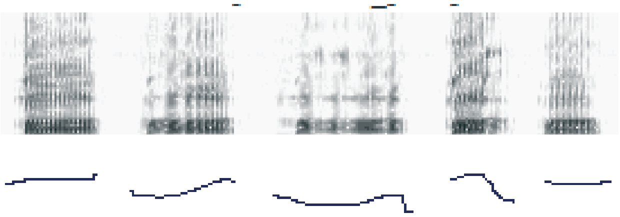

Figure 1 Spectrum (above) and F0 contour (below) of syllables /*ma1*/, /*ma2*/, /*ma3*/, /*ma4*/ and /*ma5*/

3. In addition to the 1-1 mapping between character and syllable, it is also structural inside the syllable. Each base syllable can be represented with the following form

$$\textbf{(C) + (V) V (V, N)} \qquad \text{(Form 1)}$$

According to Chinese phonology [2], first part before "+" is called *initials*, which mainly consists of consonants. There are 22 (including a zero initial). Parts after "+" are called *finals*. There are about 37 *finals* in Mandarin. Here (V), V and (V, N) are called head (glide), body (main) and tail (coda) of finals respectively. Units in brackets are optional in constructing valid syllables.

Above phonetic characteristics suggest several choices about selection of modeling unit for Mandarin speech recognition system.

- Using syllable as units. It has hardly been used for western language because of thousands of such units. However, such representation is very accurate for Mandarin and the number of units is also acceptable. However, occurrences of tri-syllable are huge and tonal syllables make the case even worse.
- Therefore, most of the current modeling strategies for Mandarin are based on the decomposition of syllable. Among them, decomposition of syllable into initial and final and some modifications become the dominant ones;
- For the tone information, there are two main ways: modeling them separately or together with base units

In the following section, we will review in detail several main strategies of modeling units about Mandarin. After fully investigating them, we propose a new modeling units design strategy, especially about tone processing in Section 3. In Section 4, we evaluate these several phone sets design strategies on both HTK and SAPI platform. Finally we summarize our results with some further discussions.

2. PREVIOUS WORK

Most modeling strategies in previous works for Mandarin LVCSR focus on the initial-final based modeling or some modifications of them as we discussed at the end of first section. One of the main differences among them is the processing of tone information. In this section, we will review several typical methods of them.

2.1. Initials plus tonal finals (ITF)

According to Form 1, previous work [3] decomposed each syllable into two parts, initial and final. According to the study of Mandarin phonology and our practice observation, most of the tone information is carried by the final part. Therefore, we assign the tone types to finals and form tonal finals. In addition, glides are assigned to finals, thus */uang2/* and */ang2/* are regarded as two different units. There are 187 units including 27 initials and 157 tonal finals in inventory (as Ph187 in Section 4).

2.2. Main vowel method

In order to reduce model size while keep model details, especially for tones, Chen et al [4] proposed a main vowel based method. They assume the tone information is exclusively carried by the main vowels of final parts instead of whole final parts. In other word, they only assign tone to body of finals (as V in Form 1). In addition, they modeled the head of finals (glides) separately. In such a way, they reduced the phone set size to about 73.

2.3. Separated tone modeling

Instead of modeling basic units and tone information together, separated modeling methods of tonal units are generally realized in two steps [5] [6]: They first model basic units without considering the tone information, then use another 5 models to deal with tones from 1 to 5 separately. During the decoding process, basic syllable candidates are generated first and then rescored with tone models to obtain the final results of tonal syllables.

3. SEGMENTAL TONAL MODELING

After studying the modeling strategies mentioned in Section 2, we can find that:

1. Separated tone modeling can realize the smallest size of phone set (about 64 units) and it was widely used for LVCSR in the early 90's. With more and more training data available and comparatively mature parameter-sharing technology, like CART based state-tying [3], such method is seldom used because of its less capability to model syllables and tones together.
2. ITF method has a good distinguishable ability among tonal syllables because it processes the finals as a whole. As a result, the co-articulation among components of finals can be ignored. However, there are still some problems about this method. First, the number of units is still quite large, especially compared with other languages (such as 40~50 for English). In addition, it can not well balance initial and tonal finals on both model size and occupied length in observations. Most of initials are far shorter than tonal finals.
3. Main vowel method has achieved some tradeoff between model accuracy and size. However, According to phonology of Mandarin, the assumption that only main vowels carry the tonal information is not true. We can also extract the meaningful F0 values from the codas for both compound finals and nasal finals. As we know, F0 represents the tone information and its contour decides the tone types.

As an alternative, we propose a new phone set construction method, called segmental tonal modeling. Basic idea consists of two parts: Assigning the glide to initial parts and forming an extended initial set in the unit inventory. The remaining parts of Form 1--v(v, n) carry the tone information. From the phonology, tone types are usually depicted with five-scale value of pitch or 2-scale (High/Low) for further simplification. The concrete relation between tone type and F0 is illustrated in Figure 2.

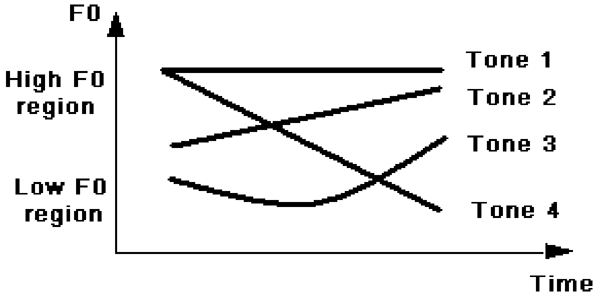

Figure 2 Illustration of Mandarin tone types vs. F0

Therefore, we model tone types by combinations of segmental 2-scale of pitch. For example, tone 1, 2, 3, and 4 are realized as high-high (HH), low-high (LH), low-low (LL) and high-low (HL) respectively. Since tone 5 is neutral one and different from other types, and we model it separately as middle-middle (MM). In addition, all finals without glides are parted into two parts and both parts carry the tone characteristics through the index of pitch with high/low/middle (or H/L/M). E.g. syllable */tiao2/* → {ti aL oH} and */huang5/* → {hu aM ngM}. More examples can be found at Table 1.

Tonal syllable	C	V1	V2
/huang1/	/hu/	/aaH/	/ngH/
/han2/	/h/	/aL/	/nnH/
/tiao3/	/ti/	/aaL/	/oL/
/da4/	/d/	/aH/	/aL/
/lou5/	/l/	/oM/	/uM/

Table 1: Segmental tones modeling strategy

Final unit inventory consist two kinds of units: extended initials, like {/ti/, /hu/}, and segmental tonal finals like {/aL/, /oH/, /aM/, /ngM/}. As we can see, /aL/ etc. is the combination of base phone (as /a/) with tone index (as L). To make it differentiate from any other tone modeling strategies, we call it segmental tones modeling.

It is natural to decompose the finals with form v(v,n) (like compound finals and nasal finals) as v and (v,n), then with proper tone index, like */ai4/*→{aH, iL} and */an2/*→{aL, nnH}. For the finals without coda (like single finals), we still part them into two segments, like */a3/*→{aL, aL}, to make it consistent with others. There are several prominent advantages for the method:

Firstly, classical tone types from tone 1 to tone 5 are modeled implicitly and jointly by segmental tones. It can reduce the size of unit inventory through better sharing among smaller units. E.g. for syllable */da/* with 5 tones, we need units {d, a1, a2, a3, a4, a5} to represent them according to ITF or main vowel methods. Now it can reduce to {d, aH, aL, aM} based on new method.

Secondly, accuracy of units is still kept through more detailed distinctions on lower layer (or smaller units) and their combinations. In addition, it is also consistent with the conclusion of Mandarin phonology on tone studies;

Thirdly, each tonal syllable is divided into three parts. When integrating with search strategies of state-of-art system, 3-part decomposition of syllable is more efficient compared with normal 2-part decomposition during the fan-out extension in search network. It is approximately twice faster as shown in next section.

Finally, assigning the glide into initial and extending original inventory of initials is meaningful. Many initials are heavily impacted by following glides due to the co-articulations. New method can increase the distinguish ability among initials through both detailed units (one unit /t/ becomes three {/t/ ti/ tu/}) and model-as-whole.

Summarily, modeling unit inventory includes 57 extended initials and 39 segmental tonal finals and totally 97 (plus *sil* model), as described as Ph97 in next section.

4. EXPERIMENTS AND ANALYSIS

We have carried out two kinds of evaluation experiments: HTK based and SAPI based. HTK based evaluation provided us the basic comparisons among different modeling strategies in pure acoustic level while SAPI based one provided us the whole performance comparison as a system according to different modeling strategies.

4.1. Evaluation based on HTK platform

HTK toolkit [7] provides us an easy way to compare acoustic models based on different modeling units. Here we tried four phone sets, called ph97, ph118, ph161 and ph187 respectively. They correspond to:

- Ph97: Segmental tonal modeling method;
- Ph118: Based on main vowels method but with extended initials same as Ph97;
- Ph161: Same as ph187 except with extended initials;
- Ph187: Initial plus tonal finals (ITF) [3];

Training set consists of 250 male speakers and about 50k utterances. Testing set consists of 25 male speakers and 500 utterances (called m-msr) as described at [3]. Error rate shown at Table 2 is base syllable error rate, a measure commonly used in Mandarin recognizer. Time shown here is based on the decoding of 20 utterances with HVite (without language model). HMM parameters size used for all configurations is nearly 6000* 8 Gaussians.

Accuracy/Speed	Four phone set configurations			
	Ph97	Ph118	Ph161	Ph187
Error rate (%)	24.79	24.15	23.61	23.95
Time (112s)	433s	538s	847s	907s
Relative time	**1.00**	1.24	1.88	2.09

Table 2: Comparison of four phone set methods based on HTK platform (m-msr)

From Table 2, we can conclude that there is no significant difference on accuracy among different methods. It is less than 5% between the best and the worst. However, the decoding speed varies greatly. Compared with former method [3], newly proposed one reduces processing time by more than 50% at the cost of less than 4% accuracy degradation. Comparison between Ph161 and Ph187 also tell us assigning the glide to initials and forming an extended initials inventory is a good choice on performance (accuracy and speed). It is suggestive for us when designing segmental tonal modeling. It also explains why we replaced original main vowel method proposed at [4] with modified one (Ph118).

4.2. Evaluation based on SAPI platform

HTK based evaluation can not show us the whole performance of new modeling strategy in system level. As an alternative, considering the tradeoff between accuracy and speed of above four methods, we apply SAPI platform, which integrated the efficient search algorithm and powerful language model, to compared ph187 and ph97 further. The details can be referred to [3]. Here a larger training corpora consisting of about 1500 speakers and about 300h wave data are used. Testing corpora consist of two sets.

- MSR: 50 speakers, 1000 utterances (about 90 min)
- CTG: 50 speakers, 1000 utterances (about 140 min)

Where MSR is more like reading speech (word perplexity of 310) and CTG is more like casual one with more speaker variability (word perplexity of 390). And test machine is a dual-CPU 730MHz with 256M memory. Error rate shown at Table 3 and Table 4 are both character error rates. Female and male each occupy half of the speakers for both testing sets.

Accuracy/Speed		Acc=0	Acc=50	Acc=100
Ph187	Error (%)	**21.037**	11.508	7.476
	Time (min)	**23.38**	40.03	71.03
Ph97	Error (%)	16.215	**10.087**	7.455
	Time(min)	15.88	**22.17**	31.57

Table 3: Comparison of two phone set methods based on SAPI platform (on testing set MSR)

Accuracy/Speed		Acc=0	Acc=50	Acc=100
Ph187	Error (%)	**29.095**	20.719	16.146
	Time (min)	**44.90**	77.98	158.13
Ph97	Error (%)	25.468	**18.931**	16.124
	Time(min)	31.52	**46.73**	72.37

Table 4: Comparison of two phone set methods based on SAPI platform (on testing set CTG)

We have tried three kinds of setups that tuning the tradeoff between speed and accuracy, called acc=0, acc=50 and acc=100 respectively. From both Table 3 and Table 4, we can observe for each setup, new phone set based on segmental tonal modeling outperform the old one in term of both accuracy and speed. Especially, for acc=0 and acc=50, the new phone set reduced the error rate by 9~23% and improve the speed by 30~45% for both testing sets. Given the similar decoding speed (as shown in bold fonts at each table), new method can reduce the error rate by 35% ~52%.

5. CONCLUSIONS AND DISCUSSIONS

A new design strategy about phone set called segmental tonal modeling for Mandarin speech recognition has been proposed in this paper. In addition to a more balanced initial inventory are formed through assigning the glides to initial, it is the first time in LVCSR, as far as we know, to model tones implicitly and jointly by segmental tones. Experimental results based on HTK and SAPI also confirmed that it is a very efficient modeling strategy. In addition to improving the accuracy by 9-23%, it greatly reduces the decoding time by more than 30~45%. Given the similar decoding speed, new phone set configuration can reduce the error rate by relatively 35%.

It is not very accurate according to phonetics studies to quantize F0 with 2-scale (H/L) and represent tone types by their combinations. Next step is to investigate the optimal combinations among segmental tones, e.g. MH instead of LH may be more accurate for the second tone, especially in spontaneous speech.

ACKNOWLEDGEMENT

The authors wish to thank Tao Chen for carrying out the HTK based experimental comparisons among different configurations on phone set.

6. REFERENCES

[1] X.D. Huang, A. Acero and H.W. Hon, "*Spoken Language Processing – a guide to theory, algorithm and system development*", pp. 428-436, Prentice Hall PTR. 2001.

[2] T. Lin and L.J. Wang, "*Phonetics Tutorials*" (in Chinese), pp. 103-121, Beijing University Press, 1992.

[3] E. Chang, J.L. Zhou, S. Di, C. Huang, and K. F. Lee, "Large Vocabulary Mandarin Speech Recognition with Different Approaches in Modeling Tones," *Proc. ICSLP'2000*, Volume II, pp. 983-986. Oct., 2000.

[4] C. J. Chen, H. P. Li , L. Q. Shen and G. K Fu, "Recognize Tone Languages Using Pitch Information on the Main Vowel of Each Syllable", *Proc. ICASSP'2001*, Salt Lake City, USA, 2001.

[5] F. Seide and N. Wang, "Two-Stream Modeling Of Mandarin Tones", *Proc. ICSLP'2000*, Beijing, 2000.

[6] L. S. Lee et al., "Golden Mandarin (I) -- A Real-Time Mandarin Speech Dictation Machine For Chinese Language With Very Large Vocabulary," *IEEE Transactions on Speech and Audio Processing*, pp. 158-179, April 1993.

[7] The HTK Toolkit: http://htk.eng.cam.ac.uk.