Acoustic Model Selection and Voice Quality Assessment for HMM-based Mandarin Speech Synthesis

Wentao Gu and *Keikichi Hirose*

Graduate School of Frontier Sciences
University of Tokyo, Japan
{wtgu, hirose}@gavo.t.u-tokyo.ac.jp

Abstract

This paper presents a preliminary study in implementing HMM-based Mandarin speech synthesis system, whose main advantage exists in generating various voices. A variety of acoustic unit representations for Mandarin are compared to select an optimal acoustic model set. Syllabic vs. sub-syllabic, context-independent vs. context-dependent, toneless vs. tonal, initial-final vs. preme-toneme models, and models with various numbers of states, are investigated respectively. To take the most advantage of HMM-based speech synthesis, some aspects affecting speaker adaptation quality, especially the selection of adaptation data size, are also studied.

1. Introduction

There have been several different acoustic models used for Mandarin speech recognition, e.g. syllable based, syllable initial-final based, and preme-toneme based approaches. In contrast, the majority of concatenation-based Mandarin speech synthesis systems employ syllables as basic acoustic units, though a few exceptions exist.

In recent years, an HMM-based approach for speech synthesis was proposed [1], where each acoustic unit was generated from HMMs instead of being selected from a corpus. In this paper, we examine the selection of acoustic units in the HMM-based approach for Mandarin speech synthesis. Several different approaches with varied settings will be compared.

Since a key advantage of HMM-based synthesis approach is adaptation capability to new voices, some factors affecting adaptation quality, such as construction of regression class tree in MLLR and selection of adaptation data size, will also be investigated.

This paper is organized as follows. The next section gives an overview on HMM-based approach for speech synthesis. Section 3 briefly describes acoustic properties for Mandarin. The data used in experiments are described in Section 4. The experimental results on selection of acoustic model will be given in Section 5. In section 6 we present the experimental results on speaker adaptation quality. Finally, Section 7 provides the conclusion of the paper.

2. HMM-based speech synthesis

In [1] an HMM-based speech synthesis approach was proposed, which employed the HMM techniques used in ASR to generate speech units directly. Its main advantage over the prevailing concatenation-based approach is the good flexibility in converting to a new speaker's voice.

The method proposed in [1] was originally implemented for Japanese speech synthesis, and was later applied to English as well in [2]. In this paper we'll implement it for Mandarin speech synthesis. As the first step, we'll focus on the modeling of spectrum, while leaving pitch and duration untouched. Therefore, pitch and duration directly extracted from the original recorded speech will be used.

The HMM-based speech synthesis system consists of 3 stages, for training, adaptation and synthesis, respectively.

In the training stage, spectral parameters are extracted from multi-speaker speech database and used to train the speaker independent (SI) HMMs. In the adaptation stage, the SI HMMs are transformed to a target speaker using speaker adaptation technique. In the synthesis stage, a sentence HMM is concatenated according to the text to be synthesized. Then spectral parameters are generated from the sentence HMM by applying the algorithm proposed in [3] (here the extracted durations are used). Finally, speech is synthesized from the generated spectrum and the extracted pitch values by using MLSA (Mel Log Spectrum Approximation) filter [4].

3. Acoustic properties for Mandarin

There are various acoustic unit representations to be chosen for Mandarin speech recognition or synthesis systems. First, since all Mandarin characters correspond to single syllables, syllable is usually regarded as a natural acoustic unit in Mandarin. In Mandarin inventory, there are about 408 base syllables (tones not attached), or around 1600 tonal syllables.

Table 1: Initial-finals and preme-tonemes in Mandarin (without tones)

	Initial / Preme	Final / Toneme
Initial -final	b,c,ch,d,f,g,h,j,k,l,m,n,p,q,r, s,sh,t,w,x,y,z,zh,ga,ge,ger, go	a,ai,an,ang,ao,e,ei, en,eng,er,i,ia,ian, iang,iao,ib,ie,if,in, ing,iong,iu,o,ong, ou,u,ua,uai,uan, ang,ui,un,uo,v, van,ve,vn
Preme- toneme	b,bi,bu,c,cu,ch,chu,d,di,du,f, fu,g,gu,h,hu,j,ji,jv,k,ku,l,li, lu,lv,m,mi,mu,n,ni,nu,nv,p, pi,pu,q,qi,qv,r,ru,s,su,sh, shu,t,ti,tu,v,w,x,xi,xv,y,z,zu, zh,zhu,ga,ge,ger,go	a,ai,an,ang,ao,e,ei, en,eng,er,i,ib,if,in, ing,o,ong,ou,u,un, U,vn

Second, there are several different decompositions of syllable. The most common way is to divide each syllable into 2 portions as initial and final. For those syllables without initial, a pseudo initial can be introduced. Syllable final is composed of an optional glide, a main vowel and an optional nasal coda.

Another kind of demi-syllabic decomposition is preme-toneme approach as proposed in [5]. Syllable initial and the glide in final (if exists) are combined into preme, while the remaining portion in final composes toneme. Therefore, preme-toneme differs from initial-final representation only when a glide is present in the syllable final.

Third, since Mandarin is a tonal language, tones in Mandarin deserve special investigation. There are 4 lexical tones and a neutral tone in Mandarin. It is generally recognized that tones in Mandarin impose on syllable final, especially concentrating on toneme portion. Whether tones are included or not will result in different set of units. As listed in Table 1 (refer to [6]), in Mandarin there are altogether 27 initials and 37 base finals, 61 premes and 22 base tonemes. When tones are taken into account, there are 157 tonal finals and 96 tonemes.

4. Experiment data

The Mandarin speech database designed by Microsoft Research Asia for recognition purpose [6] was used for training. The database consists of 19,688 sentences, with around 100 sentences for each of the 200 male speakers. It includes totally 454,291 syllable segments. Since the database was not segmented and labeled, all models were built by embedded training.

All speech signals were sampled at 16 kHz and windowed by a 25ms Blackman window with a 5ms shift. The mel-cepstral coefficients were obtained by a 30-order mel-cepstral analysis [4]. Therefore, the feature vector consisted of 93 coefficients, including the 0-th coefficients, delta and delta-delta coefficients.

Left-to-right HMMs with single diagonal Gaussian output distributions were employed (using 3-state models unless explicitly stated). We used decision tree based state clustering for context-dependent triphone modeling, unless explicitly stated. A set of phonetic questions was constructed for decision tree. For different choices of acoustic model, we revised the phonetic question list accordingly to coincide with the contextual property.

After training on the multi-speaker database, we got an SI model of average voice. For better subjective evaluation, we then adapted it using MLLR technique to a target speaker who was not included in the training set. A part of the ATR Mandarin database uttered by a female speaker [7] was used for target speaker adaptation purpose. A reason for using this database is, as designed for synthesis purpose, it is of much higher quality and hence better for synthesis and evaluation. We used 75 adaptation sentences, unless explicitly stated.

The adapted model was then used for synthesis and evaluation. In each experiment in this paper, 5 test sentences (not included in the adaptation data) from the target speaker were synthesized for 10 native Mandarin subjects to evaluate.

5. Acoustic model selection

In this section, firstly four experiments are conducted to compare (1) syllabic vs. sub-syllabic model, (2) context-independent vs. context-dependent model, (3) toneless vs. tonal model, (4) initial-final vs. preme-toneme model, respectively.

In each experiment except Experiment 2 (which will be stated separately in Section 5.2), a pair comparison listening tests was conducted for 10 subjects to pick out the preferred one from each pair of synthesized sentences. For the pair no preference could be told, it was left as *undecided.* The preference scores (in percentage) were then counted from the result of 50 pair comparisons.

At the end of this section we present the experiments to optimize the selection of number of states in syllabic and sub-syllabic models respectively.

5.1. Syllabic vs. sub-syllabic model

Different from systems for English, most concatenation-based Mandarin speech synthesis systems employ syllables instead of phonemes as the basic units, though a few exceptions exist. One reason for this is the monosyllabic nature of Mandarin characters. Another reason is that there are much fewer syllables in Mandarin inventory than in other languages such as English. In Mandarin ASR, however, most systems employ context-dependent sub-syllabic models, mainly due to the fact that there will be a huge number of context combinations when context-dependent syllabic model is considered.

Experiment 1 compared two models for HMM-based speech synthesis: one is based on tonal syllables, while the other is based on initials and tonal finals. In consideration of the data coverage difficulty for training context-dependent triphone HMMs for syllabic model, in Experiment 1 both models used context-independent HMMs for comparison.

Note that to make an equal comparison, syllabic model used 6-state HMMs, while initial-final model used 3-state HMMs. The reason for such consideration is that each syllable is composed of two demi-syllabic portions. At the same time, as will be shown in Section 5.5, such a choice happens to be optimal or sub-optimal for synthesis quality in each case.

The preference scores are shown in Table 2 (note that the remaining percentage is for *undecided* cases). The number of basic units (including silence and pause) and the number of distributions in the models are also listed. We can observe that there's no obvious preference to either side. In fact, the evaluation result varies with subjects and sentences.

However, a fairly common opinion in evaluation is that in speech generated by syllabic model some syllables are of better quality, but at the same time some conjunctions between syllables are not so continuous, like uttered in a syllable-by-syllable way. This is obviously due to the lack of cross-syllable coarticulation. On the contrary, in synthesis output of sub-syllabic model, intra-syllable artifact is more perceivable instead of any cross-syllable incontinuity.

Table 2: Comparison between syllabic model vs. sub-syllabic model

Basic unit	Preference score (%)	Num. of basic units	Num. of distributions
Tonal syllable	32	1287	7716
I-tonal F	30	186	555

5.2. Context-independent vs. context-dependent model

It's generally recognized that using context-dependent units in training HMMs will greatly improve the performance of speech recognition, if a large training database is available. Here we investigate the role of contexts in speech synthesis.

In Experiment 2, four models based on different units were compared: (1) monophone model based on initial-base finals (I-BF); (2) context-dependent triphone model based on I-BF; (3) monophone model based on initial-tonal finals (I-TF); (4) context-dependent triphone model based on I-TF.

A pair comparison listening test was conducted between each pair of models (i.e. there are totally 6 pair comparisons). In each pair, the preferred one was scored 1, while the other was scored 0. For the pair no preference could be told, both in the pair were scored 0.5. A preference score was then summed up from the 6 pair comparisons (hence its value ranged from 0 to 3). The average preference scores are listed in Table 3, where numbers of distributions are also given for comparison.

Table 3: Comparison between context-independent vs. context-dependent model

Model	mono-IBF	mono-ITF	tri-IBF	tri-ITF
Score	0.2	1.3	2.1	2.4
Num. of distrib.	195	555	2784	3408

We observe that context-dependent models consistently provide better synthesis quality than context-independent models. This is reasonable in that context-dependent model can capture coarticulation better, which is consistent with the knowledge for speech recognition. In the experiments hereafter, we will use context-dependent triphone models.

It's also shown that I-BF monophone model performs much worse than others, which is mostly due to insufficient distributions it contains. The other 3 models, improve a little progressively with the increase of number of distributions.

5.3. Toneless vs. tonal model

Another point observable from Table 3 is that tonal model is consistently better than toneless model. However, in context-independent case, tonal model shows a much higher score than toneless model, which as has been explained, is mostly due to insufficient distributions in I-BF monophone model. On the contrary, in context-dependent case, tonal model shows a score only slightly higher than toneless model.

A detailed comparison between the two models in context-dependent case was given in Experiment 3. As shown in Table 4, with almost as many as 3 times acoustic units, tonal model performs a little better than toneless model. It demonstrates that tones have a little but significant effect on spectrum quality, which is consistent with the general knowledge that the interaction between spectrum and pitch exists but is usually small.

Table 4: Comparison between toneless vs. tonal model

Basic unit	Preference score (%)	Num. of basic units	Num. of distributions
I-BF	10	66	2784
I-TF	34	186	3408

5.4. Initial-final vs. preme-toneme model

It was reported in [8] that in Mandarin speech recognition, preme-toneme (PT) model worked slightly better than or at least equally with initial-final model. In Experiment 4 we compared their roles in speech synthesis. As shown in Table 5, no significant perceptual difference was identified between synthesis outputs of the two models (both are tonal here).

For further investigation, we also did some informal listening tests to compare each speech segment separately. It was found that at some consonant-to-glide conjunctions the output of P-T model was less buzzy. This is probably due to the fact that the glide-to-vowel boundary between preme and toneme is within sonorant portion and hence smoother for perception than the consonant-to-glide (voiceless-to-voiced) boundary between initial and final. This segmental nuance, however, as shown in the evaluation result, did not affect the overall perceptual quality.

Table 5: Comparison between initial-tonal final model vs. preme-toneme model

Basic unit	Preference score (%)	Num. of basic units	Num. of distributions
I-TF	20	186	3408
P-T	18	159	3291

5.5. Selection of number of states

Here we examine the relationship between synthesis quality and the number of states in HMMs. Firstly I-TF model was investigated. We set the number of states in each HMM as 2, 3, 4, 5 and 6, respectively.

The synthesis output of 4-state model was used as the baseline for comparison, and 5-level comparison scores were identified for subjective evaluation: -2, -1, 0, 1 and 2 denote worse, a little worse, almost equal, a little better, and better quality compared with the baseline, respectively.

The average scores are listed in Table 6, where the total numbers of distributions in models are also given for comparison. It is shown that 4-state and 3-state models produce the best quality. On the one hand, 3-state is a bottom demand. On the other hand, when the number of states exceeds 4, the quality will be somewhat degraded.

Next, base syllable model (context-independent) was investigated in a similar way, except that the number of states in each HMM was set as 3, 4, 5, 6 and 7, respectively. As shown in Table 6, there's a clear tendency that synthesis quality first improves gradually and then decreases after the number of states exceeds 6. It's consistently perceived that for syllabic model 6-state HMMs work the best.

Table 6: Comparison of I-TF model and base syllable model with different number of states

Num of states		2	3	4	5	6	7
I-TF model	Score	-2.0	**-0.1**	**0**	-0.6	-0.5	/
	num of dist	2558	3408	3902	4293	4673	/
base-S model	Score	/	-2.0	-0.8	0	**0.7**	0.1
	num of dist	/	1215	1620	2025	2430	2835

6. Speaker adaptation quality

A critical advantage of HMM-based speech synthesis is its big convenience to convert to a new speaker by adapting the HMM parameters with a small amount of data. In such a case where very few adaptation data is available, MLLR adaptation approach is a good choice. In this paper we only adapted mean vectors with a block diagonal transformation matrix.

Context-dependent I-TF model was used for experiments here. Firstly some informal listening tests were taken on the synthesis output with 75 adaptation sentences. By informal listening, we found that a global adaptation was not adequate to produce a good voice quality. Therefore, a regression class tree was constructed to pool the transforms automatically.

It was perceived that the adaptation quality was significantly improved when 2 base classes were involved (basically divided into voiceless consonants and vowels or sonorants). However, when we continued to refine the tree classification by increasing the number of base classes from 3 to 15 progressively, further voice improvement was quite trivial. In the following experiment, we employed regression class tree with 6 base classes.

To investigate the relationship between adaptation quality and the size of adaptation data, we adapted the model to the target speaker by using 2, 4, 6, 8, 10, 12, 15 and 75 adaptation sentences respectively. Synthesis output with 75 adaptation sentences was used as the baseline for comparison, and 4-level comparison scores were identified for subjective evaluation: 0 (almost equal), -1 (a little degraded), -2 (degraded but still acceptable), and -3 (obviously worse).

The average scores are shown in Figure 1, from which we can observe that 8 sentences (with 379 syllables) will give an acceptable quality, while 12 ~ 15 sentences (with 579 ~ 670 syllables) will produce the quality almost comparable with that adapted from a much bigger database. It turns out that it's even not necessary to involve a large adaptation database when MLLR technique is employed.

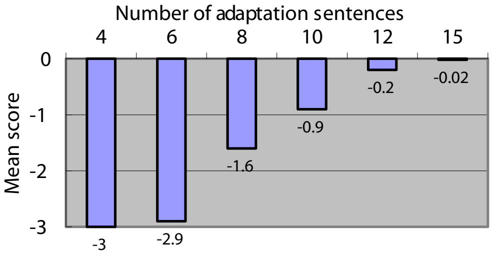

Figure 1: Evaluation of adaptation quality with different number of adaptation sentences

7. Conclusion

We have compared various selections of acoustic model for HMM-based Mandarin speech synthesis. The experiments show that in context-independent case, there is no clear preference between syllabic model and sub-syllabic model. For sub-syllabic model, however, using context-dependent units will improve the synthesis voice quality. Among sub-syllabic approaches, preme-toneme model tends to produce less buzzy consonant-to-glide conjunctions than initial-final model, which however gives negligible perceptual difference. Besides, we notice that in spectrum modeling the employment of tonal acoustic results in a little but significant quality improvement.

In conclusion, based on our speech database, the context-dependent tonal sub-syllabic model with 4 or 3 states performs the best for spectrum modeling in HMM-based Mandarin speech synthesis. In specific tasks, however, the selection of acoustic model can be made a trade-off according to the scale of available training database.

The factors affecting speaker adaptation quality in HMM-based Mandarin speech synthesis are especially investigated. On the one hand, at least 2 base classes should be included in the regression class tree for MLLR adaptation, while further classification results in little improvement. On the other hand, 8 ~ 15 sentences (i.e. around 400 ~ 600 syllables) will generally be adequate to achieve a good adaptation quality.

In the future work, we'll re-examine these issues when modeling the spectrum, pitch and duration simultaneously.

8. Acknowledgements

The authors would like to thank Microsoft Research Asia and ATR for their kindly providing the Mandarin speech database.

9. References

[1] Tamura, M., Masuko, T., Tokuda, K., etc., "Text-to-speech synthesis with arbitrary speaker's voice from average voice", *EuroSpeech-2001*, pp.345-348, 2001.

[2] Tokuda, K., Zen, H., and Black, A. W., "An HMM-based speech synthesis system applied to English", *IEEE 2002 Workshop on Speech Synthesis*, 2002.

[3] Tokuda, K., Kobayashi T., and Imai, S., "Speech parameter generation from HMM using dynamic features", *ICASSP-95*, pp.660-663, 1995.

[4] Fukada, T., Tokuda, K., Kobayashi, T., and Imai, S., "An adaptive algorithm for mel-cepstral analysis of speech", *ICASSP-92*, pp.137-140, 1992.

[5] Chen, C. J., Gopinath, R. A., Monkowski, M. D., etc., "New methods in continuous Mandarin recognition", *EuroSpeech-97*, pp.1543-1546, 1997.

[6] Chang, E., Shi, Y., Zhou J., and Huang, C., "Speech lab in a box: a Mandarin speech toolbox to jumpstart speech related research", *EuroSpeech-2001*, pp.2799-2802, 2001.

[7] Ni, J., Kawai, H., "A synthesis-oriented Mandarin speech corpus", *Autumn Meeting of the Acoustic Society of Japan*, pp.319-320, 2002.

[8] Lo, W. K., Meng, H. M., and Ching, P. C., "Sub-syllabic acoustic modeling across Chinese dialects", *The Second International Symposium on Chinese Spoken Language Processing*, pp.97-100, 2000.