A MANDARIN TTS SYSTEM WITH AN INTEGRATED PROSODIC MODEL

ShaoHuang Pin[1], Yehlin Lee[1], Yong-cheng Chen[2], Hsin-min Wang[2], and Chiu-yu Tseng[1**]*

[1]Phonetics Lab, Institute of Linguistics, Academia Sinica, Taipei
[2]Institute of Information Science, Academia Sinica, Taipei
E-mail: *whm@iis.sinica.edu.tw,**cytling@sinica.edu.tw

ABSTRACT

Phrase grouping is essential to characterize the prosody for Mandarin fluent speech. Evidence of prosodic phrase grouping has been found both in adjustments of F_0 contours and temporal allocations within and across phrases. In this paper, we discuss the development of a Mandarin TTS system that integrates the prosody processing modules, such as duration modeling, F_0 modeling, and break predictions. The database consists of 1292*3 syllable-tokens chopped off specially designed three-phrase carrier sentences. Since temporal allocations and rhythmic structure in speech flow are carefully dealt with, the TTS system is capable of converting long paragraph text input into natural synthesized speech output.

1. INTRODUCTION

In modern concatenative text-to-speech (TTS) systems, there are two completely different trends in the size of the speech database. One is to use a small database that only consists of the basic set of phonemes. This kind of TTS system can reach basic intelligibility, but usually lose naturalness. The other is to use a huge database that covers as many context variations as possible. This kind of TTS system may reach acceptable naturalness, but lose freedom of mobility. Since a speech database of unlimited size is infeasible and a huge database still contains limited context variations only, a TTS system will never output natural speech as a human being without understanding the prosodic organization of human speech. It is even more challenging to maintain the naturalness under the constraints of limited database size. The prosodic model, synthesizing tool, and specially designed database are all crucial to the development of a good TTS system.

In our previous studies, we showed that phrase grouping is essential to characterize the prosody for Mandarin fluent speech [1]. This model is based on the unit located inside the perceived different levels of boundary breaks across speech flow. The boundaries are marked using the ToBI-based self-designed labeling system [2] that tagged small to large boundaries with a set of break indices (BI); i.e., B1 to B5. Evidence of prosodic phrase grouping was found both in adjustments of F_0 contours and temporal allocations within and across phrases [1, 2]. Under this framework, phrases are no longer unrelated intonation units. This framework can also be viewed as a tree-branching hierarchical organization. From top down, the layered nodes are phrase groups (PG), breath groups (BG), prosodic phrases (PPh), prosodic words (PW), and syllables (SYL). These constituents are, respectively, associated with break indices B5 to B1. We have conducted the above prosody analysis on a female-read speech database of 26 long paragraphs or discourses in text, or a total 11592 syllables (or Chinese characters). The speech data were manually labeled by trained transcribers for perceived prosodic boundaries (BI) and aligned with transcripts. A duration model, a F_0 model, and a break prediction model have been trained from this database. These models can be used by the text processing module of the TTS system to produce the prosodic structure of any given text.

Since the Chinese writing system consists of mono-syllabic logographic characters, and since there are only 1292 distinct tonal syllables, it is reasonable to choose syllables as the concatenative units. Considering the phenomenon of prosodic phrase grouping, we specially designed a three-phrase carrier sentence to record the syllable tokens. Each syllable was embedded in the carrier sentence at the initial, medial, and final positions, respectively. Therefore, the database is made of 1292*3 Mandarin tonal syllable tokens.

Based on the prosody processing modules, such as duration modeling [3], F_0 modeling [4], and break predictions [5], we have implemented a Mandarin TTS system using the syllable-token database. Since the hierarchical PG structure of fluent speech was adopted as the bases of prosody modeling, the TTS system is capable of converting long paragraph text input into natural synthesized speech output.

2. DURATION MODELING

Our duration model of the rhythmic patterns in Mandarin speech flow [3] reveals that the syllable duration is not only affected by the syllable constitution itself, but also affected by the upper layer prosodic structures, namely PW, PPh, BG, and PG, respectively.

The corpus used for training the duration model was female read speech data of 26 long paragraphs or discourses in text, or a total 11592 syllables (or Chinese characters). The speech data were first automatically aligned with initial and final phones using the HTK toolkit, and then manually labeled by trained transcribers for perceived prosodic boundaries or break indices (BI).

2.1. Intrinsic statistics of syllable duration

A layered model is used to estimate the syllable's duration. At the SYL-layer, a linear model is adopted,

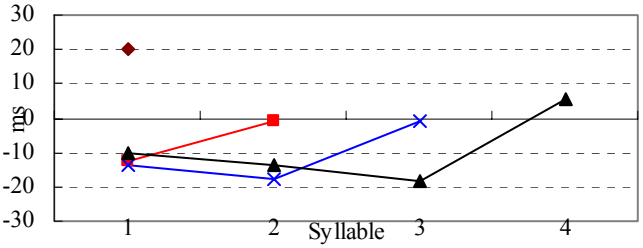

Figure 1. *Rhythmic patterns in PW-layer*

Syllable intrinsic duration = constant + CTy + VTy + Ton
+ PCTy + PVTy + PTon + FCTy + FVTy + FTon
+ 2-way factors of the above factor
+ 3-way factors of the above factor (1)

The *constant* was set to 185 ms, which was calculated from the corpus. *CTy*, *VTy*, and *Ton* represent the offset values corresponding to the consonant type, vowel type and tone of the current syllable, respectively. Prefix of *P* and *F* represent the corresponding factors of the preceding and following syllable. The 2-way factors consider the joint effect of two single-type factors. There are C_2^9 (=36) 2-way factors in total. The 3-way factors consider the joint effect of three single-type factors. The 3-way factors with a negligible influence on the syllable duration were excluded from consideration. Only three 3-way factors were left, they are the combination of consonant type, vowel type and tone of the preceding, current, and following syllables, respectively. As a result, a total of 49 factors were considered. The 21 consonants and 39 vowels (including diphthongs) of Mandarin were, respectively, grouped into 7 and 9 categories according to their measured mean duration. Notice that the SYL-layer model is independent of the prosodic structure. The SYL-layer model can explain about 60% of syllable duration.

2.2. The effect of layered prosodic structure

As depicted in Figure 1, the syllable duration is affected by its position within a PW. Note that the PW final syllable tends to be lengthened compared to other syllables. The residue error that can not be explained at the SYL-layer can be further explained by the PW-layer. Accordingly, the syllable duration is postulated as:

DurS (ms) = Syllable intrinsic duration
+ f_{PW}(PW length, position in PW). (2)

Since the syllable intrinsic duration is the duration controlled by the SYL-layer, the PW-layer has its effect of speeding the rhythm by subtracting a value derived from Figure 1 and vice versa.

The PPh-layer affects the syllable duration in a similar way as the PW-layer. As to the BG-layer or above, the length of the prosodic unit gets longer and complicated, the perceived significance exists only in the initial and final PPh units. Therefore, we model BG-layer's effect as the effect in the initial and final PPhs in the BG-layer. The overall model is thus formulated as:

DurS (ms) = Syllable intrinsic duration
+ f_{PW}(PW length, position in PW)
+ f_{PPh}(PPh length, position in PPh)
+ f_{IFPPh}(Initial/Final PPh length, position in PPh) (3)

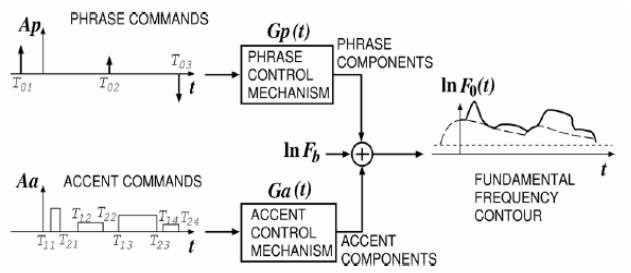

Figure 2. *Generation process of the Fujisaki model (from Fujisaki, 1982)*

2.3. Application to the TTS system

The results of these duration patterns are not only evidence of interaction between syllable duration adjustment and prosodic level units, but also a useful duration prediction method. Therefore, temporal allocation is implemented in our TTS system. The details will be described in Section 4.2.

3. F_0 MODELING

There are many F_0 models around the world. We use the well-known Fujisaki model as the production model of F_0 [6]. Figure 2 shows the basic structure for producing the F_0 contour. The model connects the movements of cricoid's cartilage to the measurements of F_0 and is hence based on constraints of human physiology. Therefore, it is reasonable to assume that the model could accommodate F_0 output of different languages. Successful applications of the model on many language platforms have been reported, including Mandarin [7, 8].

In the case of Mandarin Chinese, phrase commands were used to produce intonation at the phrase level while accent commands were used to predict lexical tones at the syllable level [9]. Phrasal intonations are superimposed on sequences of lexical tones. Therefore, interactions between the two layers cause modifications of F_0 to produce the final output. The superimposing of a higher level onto a lower level leaves room for even higher level(s) of F_0 specification to be superimposed and built. Thus, we decided to implement our PG framework of phrase/intonation-grouping on the Fujisaki model by adding a PG layer over phrases. In other words, after generating phrasal intonations for each phrase, PG specifications were then superimposed onto phrase strings subsequently. By adding one higher level of PG specification, the F_0 patterns of phrase grouping could be achieved.

3.1. Building the phrasal intonation model

The corpus used here is the same as the one used for duration modeling. We first proceeded with automatic parameter extraction, and then used the extraction results to build a statistical phrasal intonation model. Again, a linear model is adopted:

Phrase command Ap = constant + coeff1 × pause
+ coeff2 × pre_phr + coeff3 × $f_0$min + iSyl_PPh, (4)

where pause is the speechless portion at the beginning of a phrase command, *pre_phr* is the accumulated magnitude of previous phrase commands when the response of the current

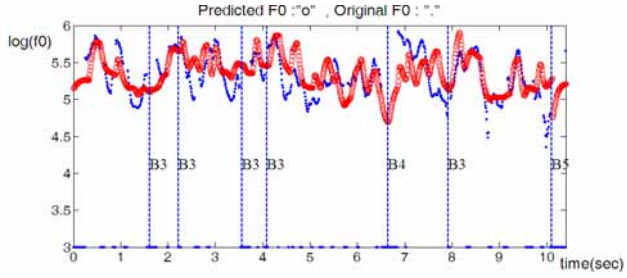

Figure 3. *Simulation result of global intonation modeling of a PG. The red line (circle) represents predicted global contours; the blue (dot) represents contours of the original speech F_0 data.*

phrase command reaches to the peak, f_0min is the minimum fundamental frequency of the utterance, and *iSyl_PPh* is the index of the syllable in PPh where the related phrase command is located. The accumulated magnitude of previous phrase commands at time *t* can be calculated as:

$$\sum_i A_{pi} \cdot \alpha^2 \cdot (t - T_{0i}) \cdot e^{(-\alpha(t-T_{0i}))} \quad (5)$$

where *i* is the index of previous phrase commands, while A_{pi} and T_{0i} are the magnitude and timing, respectively.

3.2. Building the PG intonation model

As shown in [10], the PG intonation has significant effects in the first and last PPh units only. Therefore, the intonation model can be modified as:

Phrase command Ap = constant + coeff1 ⨯ *pause*
+ *coeff2* ⨯ *pre_phr + coeff3* ⨯ f_0min
+ *iSyl_PPh* (syllable position in PPh)
+ *PG effect coefficients* (Initial/Final PPhs) (6)

Figure 3 shows a comparison between an F_0 prediction/production of a PG and the original intonation.

3.3. Application to the TTS system

The constructed PG intonation model can be applied to our Mandarin TTS system to produce F_0 contours. Since the higher level of prosodic unit is taken into account, more fluent and natural intonation can be obtained. The details of adjusting the F_0 output will be described in Section 4.3.

4. THE TTS SYSTEM

4.1. Speech database

Both the duration and F_0 models described above are built based on the PG structure. Therefore, we have specially designed our database such that the TTS system can be implemented to use these models.

The database is made of 1292*3 Mandarin tonal syllable tokens. Each of the 1292 syllables was embedded in a phrase of a three-phrase carrier sentence (a PG of 3 PPhs) in initial, medial, and final positions, respectively. The speech data were recorded by a native female speaker in a sound-proof room. The target syllable tokens were listened to and manually edited from the carrier sentence by trained transcribers. Figure 4 shows the

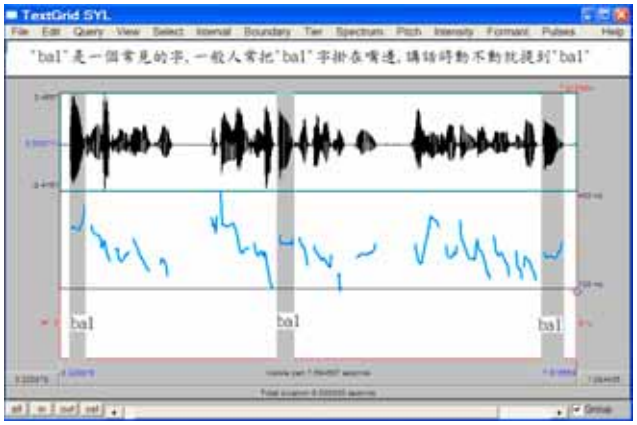

Figure 4. *The carrier sentence with target syllable "ba1" embedded in it. ("ba1" is a frequently used syllable, people say"ba1" very often, often times when people speak, they'd use "ba1").*

carrier sentence with syllable "ba1" embedded in it, the associated wave, the F_0 contour, and the time stamps of the target syllable "ba1". In our TTS system, the time-domain pitch-synchronous overlap-add (TD-PSOLA) [11] method is employed to perform prosody modification. The pitch marks were first automatically estimated, and then manually repaired by trained transcribers.

For each syllable, there are 3 tokens, respectively, collected from the initial, medial, and final positions of a PG. Since the prosodic models were trained by a different speaker's speech, the models need to be adapted to satisfy the condition indicated by the initial, medial, and final syllables of a PG to be synthesized. In other words, the TTS system will only adjust the duration and F_0 of the other syllables using the modified prosodic models but keep those of these 3 syllables unchanged.

4.2. Duration adjustment

Since the TTS database was from a different speaker, the absolute duration predicted by the duration model should be adjusted, while the rhythmic patterns in the PG organization should be kept.

Because the initial, medial, and final syllables are originally collected from the same positions of a PG, their durations should not be changed. The durations of the rest syllables, which were originally the first syllable of a PW at the medial position of a medial PPh of a 3-PPh PG, should be modified to satisfy the rhythmic pattern in the PG organization. In this way, to synthesis a PG of *m* characters (or syllables), the duration of the *i*-th syllable is given by

$$DurS_i^* = \begin{cases} OriDur(S_i) & ,i = 1, m/2, m \\ OriDur(S_i) - DF_i & ,1 < i < m/2, m/2 < i < m, \end{cases} \quad (7)$$

where $OriDur(S_i)$ is the corresponding syllable-token's original duration and DF_i is an offset factor, which is calculated by

$$\begin{aligned} DF_i &= M_{TC} / M_{MC} \times [f_{PW}(PW\ length, position\ in\ PW) - f_{PW}(2,1) \\ &+ f_{PPh}(PPh\ length, position\ in\ PPh) - f_{PPh}(11,6) \\ &+ f_{IFPPh}(Initial / Final\ PPh\ length, position\ in\ PPh)], \end{aligned} \quad (8)$$

where M_{TC} and M_{MC} are, respectively, the mean of syllable duration of the TTS corpus and the training corpus, and $f_{PW}(\ )$, $f_{PPh}(\ )$ and $f_{IFPPh}(\ )$ are the same as that in Eq. (3), which were calculated from the training corpus.

4.3. F_0 adjustment

In the PG intonation model described in Section 3, the output of the model can be shifted up or down in order to adjust the F_0 level of the initial syllable. Based on the characteristics of the Fujisaki model's phrase commands, the magnitude *Ap* could be altered to fit a new peak level. If the predicted intonation peak *P* has to be changed to a new peak level P', the original *Ap* should be altered to the new Ap'. The relationship between *P* and *Ap* [12] is governed by,

$$Ap \cong 2.718 \times P / \alpha , \quad (9)$$

where *P* is the peak of the response of the phrase command with magnitude *Ap*, and α is the time constant of the phrase command.

Once we know the peak level, we can derive the corresponding magnitude of the phrase command. In the adaptation scheme of our intonation model, we first predict all the necessary phrase commands. Then, we adjust the first peak according to the F_0 of the first syllable selected from the TTS database by the method described above. Finally, we adjust the rest of phrase commands with the same percentage as applied to the first phrase command. In this way, we can adapt to the intonation level of the TTS database but keep the overall intonation contour shape.

4.4. Break prediction

The prosodic boundaries and break indices are predicted by analyzing the syntactic structure of the text to be synthesized. The details are described in the other paper submitted to the same conference [5].

4.5. System flowchart

Given a text, first of all, the prosodic boundaries and break indices will be predicted based on the analysis of syntactic structure. The PG hierarchical structure and the pronunciations (the syllable sequence associated with the text) will be generated as well. Then, the durations of all syllables will be assigned by the duration model, while the F_0 contours of all phrases will be generated by the intonation model. All the outputs of text processing will be stored in a predefined XML document. Finally, the TD-PSOLA method is employed to perform prosody modification, and the TTS system will output the concatenative waveform.

5. DISCUSSIONS

Our TTS system aims at synthesizing fluent speech in long paragraphs. Because long speech paragraphs are perceived with its significant initial and final PPhs, modeling this phenomenon will signal output topics clearer in multiple phrases groups. The duration model was clear in each layer, thus a straightforward linear model was sufficient to model durational effect of every prosodic unit. The F_0 model based on the Fujisaki model is more complicated but we used the extensible ability of the Fujisaki model to extend the F_0 model to the overall intonation of PG.

6. CONCLUSIONS

We believe that an integrated prosodic model that organizes phrase groups into related prosodic units to form speech paragraphs will significantly improve output naturalness for unlimited TTS. The speech database also illustrates how mono-syllables could be collected to offer more prosody information. Future work includes building the TTS system on larger amount of more varied speech data, and incorporating more prosodic information as well.

7. REFERENCES

[1] C. Tseng, S. Pin, and Y. Lee, "Speech Prosody: Issues, Approaches and Implications", in Fant, G., H. Fujisaki, J. Cao and Y. Xu Eds. *From Traditional Phonology to Mandarin Speech Processing,* Foreign Language Teaching and Research Process, Beijing, China, 2004, pp. 417-438.

[2] C. Tseng, and F. Chou, "A Prosodic Labeling System for Mandarin Speech Database", *Proceedings of ICPhS99*, pp. 2379-238.

[3] C. Tseng, and Y. Lee, "Speech Rate and Prosody Units: Evidence of Interaction from Mandarin Chinese", *Proceedings of Speech Prosody 2004*, pp. 251-254.

[4] C. Tseng, and S. Pin, "Mandarin Chinese Prosodic Phrase Grouping and Modeling - Method and Implications", *Proceedings of International Symposium on Tonal Aspects of Languages—with Emphasis on Tonal Languages (TAL 2004)*, pp. 193-19.

[5] K. Chen, C. Tseng, H. Peng, and C. Chen, "Predicting Prosodic Words from Lexical Words - A First Step towards Predicting Prosody from Text", submitted to *International Symposium on Chinese Spoken Language Processing*, 2004.

[6] H. Fujisaki, and K. Hirose, "Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese", *Journal of the Acoustical Society of Japan (E),* 5(4): pp. 233-241, 1984.

[7] H. Fujisaki, "Modeling in the Study of Tonal Feature of Speech with Application to Multilingual Speech Synthesis", *Proceedings of SNLP-O-COCOSDA 2002*.

[8] H. Mixdorff, "Quantitative Tone and Intonation Modeling across Languages", *Proceedings of International Symposium on Tonal Aspects of Languages- with Emphasis on Tone Languages (TAL 2004)*, pp. 137-142.

[9] H. Mixdorff, Y. Hu, and G. Chen, "Towards the Automatic Extraction of Fujisaki Model Parameters for Mandarin", *Proceedings of Eurospeech 2003*.

[10] C. Tseng, and S. Pin, "Modeling Prosody of Mandarin Chinese Fluent Speech via Phrase Grouping", submitted to *ICSLT-O-COCOSDA 2004*.

[11] M.J. Charpentier, and M.G. Stella, "Diphone Synthesis using an Overlap-Add Technique for Speech Waveforms Concatenation", *Proceeding of ICASSP86*, pp. 2015-2018.

[12] H. Mixdorff, "A Novel Approach to the Fully Automatic Extraction of Fujisaki Model Parameters", *Proceedings of ICASSP2000*, pp. 1281-1284.