CORPUS-BASED MANDARIN SPEECH SYNTHESIS WITH CONTEXTUAL SYLLABIC UNITS BASED ON PHONETIC PROPERTIES

Fu-chiang Chou[1] *and Chiu-yu Tseng*[2]
[1] Department of Electrical Engineering, National Taiwan University
[2] Institute of Linguistics, Academia Sinica
Taipei, Taiwan, Republic of China
email-addr: moza@speech.ihp.sinica.edu.tw

ABSTRACT

This paper describes an improved concatenative synthesis module for a Chinese text-to-speech system [1]. The concatenated segments are on-line selected from a designed speech corpus that is precisely segmented with an improved version of HMM models. The selection criteria are the prosodic and contextual similarities between the units and the desire targets from the previous module of the TTS system. The TD-PSOLA modifies the prosodic parameters of the selected units, and three methods for unit concatenation are performed according to the types of the syllabic junctures. These types are classified with the knowledge from the phonetic observations of large amounts of speech data. The output speech is remarkably fluent and natural because the coarticulation effects cross syllabic boundaries are well modeled and less prosodic modification is needed for the TD-PSOLA.

1. INTRODUCTION

Conventional speech synthesis methods use a limited inventory of phonemes, diphones, or demisyllables as basic synthesis units. For Mandarin, the majority of the synthesizers are based on syllabic units because there are about only 400 syllables without tones in Mandarin Chinese[2][3]. But this kind of units does not allow one to model the coarticulation effects that cross syllable boundaries. As a result, short silences must be inserted between adjoining syllables to avoid discontinuity if the syllables are directly concatenated. This procedure results in an unnatural 'choppiness' [4]. The other shortage is that only one instance of unit is modified to fit into the different prosodic and contextual environments. But we know that high quality synthesis depends both on appropriate segmental and prosodic properties for each source unit. Therefore, as the capability to handle large amounts of corpus increase, all segments of a speech database can be used as the synthesis units. This is so-called corpus-based synthesis. With such an approach, the main factors that influence the output quality are:

1. Corpus: Given an infinite speech corpus and an efficient index, we can produce synthetic speech that is almost indistinguishable from human speech. But with a medium-size corpus, we must carefully design the contents in order to cover most of the speech units and variations in that language. Furthermore, the corpus must be correctly labeled with segmental and prosodic information in an efficient way.

2. Units: The adequate size of the synthesis units is language dependent. Except the fixed-size units such as diphones, triphones or demisyllables, non-uniform units are also investigated in [5][6]. The major problem is how to select the suitable units -- the criteria and the methods. Except the segmental properties, more attentions are paid to the prosodic similarity in recent research [7].

3. Concatenative methods: Segments can be directly concatenated or overlap with smooth transition. Concatenated points can be the unit boundaries or inside a stable part of the units. In the research of the optimal coupling technique, the concatenated points are on-line chosen to provide minimum mismatch with the neighboring units [8]. These variations of concatenative methods indeed depend on the phonetic properties of the two concatenated units.

With considering the above factors, we construct a corpus-based Mandarin speech synthesis system with contextual syllabic units based on phonetic properties. The corpus is designed to cover most of the syllabic units and combinations. The segmental alignment and prosodic labeling are performed automatically with little errors. Syllables are the basic units but additional prefix and suffix parts are included to solve the coarticulation problems. These units are selected to match the segmental and prosodic properties with the desired targets, and they are further concatenated with three methods according to the phonetic types of the syllabic junctures. The end purpose of the whole system is to select the best segments from the speech database and generate the most natural-sounding speech output.

This paper is organized as follows. In Section 2 we discuss the design and labeling of the speech corpus. In

Section 3 we then describe how to select and concatenate the speech segments. Finally we summarize the major findings and future works.

2. CORPUS DESIGN AND LABELING

2.1 Corpus Design

In Mandarin, there are about 400 syllables without tones, which means that there are 16,000 pairs of di-syllables. It's hard to cover all these combinations in a speech corpus. In order to describe transitions between syllables, units smaller than the syllable must be chosen. An INITIAL/FINAL format can describe the composition of a Mandarin syllable. INITIAL is the initial consonant and FINAL is the vowel (or diphthong) part with an optional medial or a nasal ending. In theory, there are about 2000 FINAL-INITIAL and FINAL-FINAL (No INITIAL in the latter syllable) patterns in the disyllabic junctures. The speech corpus is designed to cover most of these combinations [9]. Moreover, the corpus is organized as many short paragraphs so as to cover many prosodic variations. Six professional speakers read the corpus at a normal speaking rate. If there are hesitations or mistakes, the speaker will be asked to read the sentence again until each character is correctly pronounced. This can reduce the errors for further segmentation and labeling.

2.2 Automatic Segmentation

With an accompanying orthographic transcription, the corpus can be segmented by labeling with the HMMs [10]. The units of the HMMs are context independent INITIALs and FINALs trained by the HTK toolkit. The feature vectors include 12 dimensions of MFCC, 1 dimension of RMS power and their differential values. The frame rate is set to 5ms to increase the precision of the segmentation. The results showed that if the HMMs are not trained by the data of the same speaker with manual labels, the performance is always not satisfactory and required further manual adjustments. We call these processes a semi-automated segmentation method.

During the manual processing of the speech corpus, we found that most of the errors can be classified and adjusted with some rules. The errors are classified with the combinations of phonetic types, for examples: nasal+vowel, vowel+fricativc, etc. These rules can be implemented into an algorithm to post-process the label files and save the manual efforts. The adjusted results are applied to adapt the parameters of the HMMs. The block diagrams of the segmentation system are illustrated on Fig. 1. The input is the speech signal and its syllabic transcription; the output is the INITIAL/FINAL sequence with the associated position.

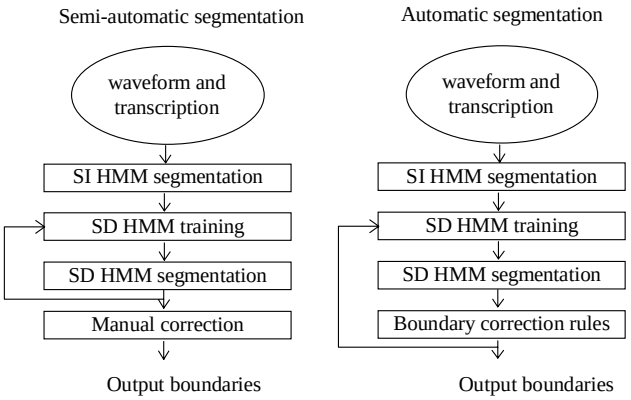

Figure 1. Block diagrams of the two kinds of segmentation process (semi-automatic and automatic).

For semi-automatic processing, the SI (speaker independent) HMMs performs a rough segmentation for initial training of the speaker dependent HMMs. The parameters of the SD HMMs update with an iterative training process. The outputs of the final segmentation from HMMs are then adjusted with human experts.

For automatic processing, the boundary correction rules are applied instead of the human correction. These prior described rules are based on the knowledge from the observations in human correction procedures. The outputs of SD HMMs are accepted as the initial boundaries. The program then searches in a local area for the acoustic features that match the phonetic properties of the units. The features include RMS power, voicing probability and FFT spectrogram derived from ESPS programs. The window sizes are varied from 5ms to 20ms according to the features and phonetic types of units. For example, a 5ms window of RMS power is applied to locate a plosive because there is a short burst of energy when the sound is released. If the specified acoustic features are not found in that area, the boundary is left no change. The adjusted boundaries are further processed to update the parameters of the SD HMMs. These procedures are recursively performed until the average alternation of boundaries is under a threshold.

To evaluate the effects of the whole process, the output after the manual correction is set as the reference. The errors are calculated as the difference between the determined boundaries and the reference boundaries. The segmentation rate is defined as the percentage of errors within 10ms and 20ms. Without the boundary correction rules, the mean error of the HMMs is 14.2ms, and the segmentation rate is 66.3% (91.2%) within 10ms (20ms). By retraining the HMMs with the boundary correction rules, the average error of the outputs decreases to 8.3ms, and the segmentation rate within 10ms (20ms) increases to 78.4% (96.5%).

2.3 Prosodic and Contextual Labeling

The prosodic and contextual parameters for each syllable in the speech corpus are automatically labeled after the decision of the syllabic boundaries. The prosodic parameters are mainly derived from the output of the ESPS get_f0 program, and the contextual parameters are the phonetic types of the segments preceding and following the syllable. All information listed below is saved in an index file for quick retrieval.

- syllable ID (preceding, current and following)
- boundary position (start and end)
- possible connection position (INITIAL and FINAL)
- duration (INITIAL, FINAL and syllable)
- fundamental frequency (8 points cross the syllable)
- RMS power (head, middle, tail and average)

3. UNIT SELECTION AND CONCATENATION

3.1 Classification of Disyllabic Juncture

To effectively utilize the speech corpus, the types of the disyllabic junctures are first examined. Based on the classification of the FINAL/INITIAL patterns, the disyllabic junctures fall into 3 types as listed in Table 1.

Types of disyllabic junctures		Class of INITIAL for latter syllable				
		plosive & Affricate	fricative	nasal	lateral	no INITIAL (vowel)
Class of FINAL ending for former syllable	vowel	1	3	2	3	3
	nasal	1	2	2	2	2

Table 1. The types of disyllabic juncture in Mandarin Chinese

Type 1: When the INITIAL of the latter syllable is a plosive or an affricate, there will be a short period of silence and a clear spectral change during the juncture. The boundary can be easily detected and with no obvious coarticulation effects.

Type 2: There is a nasal ending in the FINAL of the former syllable or a nasal INITIAL in the latter syllable. The adjoined phoneme could be affected by the nasal, but the spectral properties during the nasal are extremely stable.

Type 3: In the other cases, the spectral properties are smoothly transformed from the FINAL of the former syllable to the INITIAL of the latter syllable. Most of the coarticulation effects occurred in these situations.

3.2 Concatenative Methods

According to the three types of the syllabic junctures, there are three ways to concatenate two units.

1. Hard concatenation: The simplest way to put two units together and no smoothing is needed. For example, the concatenated point is during the silence period for the type 1 juncture.

2. Diphone concatenation: Two units are concatenated within a region that have the same spectral properties. A suitable segment for diphone concatenation must have a stable part that is insensitive to contextual influences. Type 2 juncture is a good example because the nasal ending or the initial nasal is relatively quite stable.

3. Soft concatenation: The concatenation takes place at the syllabic boundaries. However, the segments are smoothed by including the transition parts that appear in the speech database. A certain amount of overlap between the two units makes the transition smoothly. This is suitable for Type 3 juncture.

3.3 Units Selection

The input for the synthesis module is a syllable string with correspondent F0 contour, duration and energy values. These values are generated from the previous module of the TTS system. To output natural speech, the suitable units must be selected and concatenated before further prosodic modifications. At first, the syllabic units that fit the syllable string are selected from the index file, and then form a lattice of the syllabic units. The best sequence is then determined by the prosodic modification cost and contextual mismatching cost with a Viterbi decoding process.

The prosodic modification cost $Cp(u_i,t_i)$ is defined as the normalized distance of the prosodic features between the selected unit(u_i) and desired target(t_i).

The contextual mismatching cost $Cc(u_i,u_{i+1})$ is the cost defined to model the contextual mismatching effect when the two units are concatenated. The average spectral distances between two units with different context are calculated in advance and stored in a table so the on-line computation load can be decreased. The best unit sequence is the path that minimizes:

$$\sum_{j=0}^{n} (W_p \bullet Cp(u_j, t_j) + W_c \bullet C_c(u_j, u_{j+1}))$$

where W_p and W_c are the prosodic weighting and contextual weighting. Currently, these weights are hand-tuned with informal subjective listening. Fig. 2 illustrates a hypothetical example.

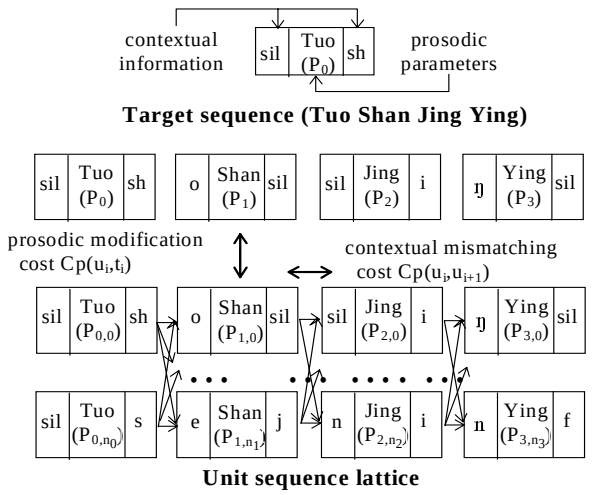

Figure 2. A hypothetical example of a syllabic lattice for units selection.

After the decision of the concatenated units, TD-PSOLA is performed to modify the prosodic features of the selected units. The modified units are concatenated with the three methods mentioned in section 3.2 and output to the speakers. Due to time constraints, no formal assessment method is designed to evaluate the quality of the synthetic speech yet. Informal test confirms that the speech is more natural and fluent than simple concatenation of the syllabic units. But occasionally the quality degradation is caused by the segmentation errors or pitch marker errors. This is an unavoidable problem in the operation of large amounts of speech data. How to delete the segments with unreliable boundaries and pitch markers is under further investigations.

4. SUMMARY

There is much progress in the technology of speech synthesis in recent years. Due to the research of speech recognition, large speech corpus can be automatically segmented. Due to the increasing of the computer memory and computation power, the synthesis units can be optimally selected on-line. To integrate these technologies with the special phonetic properties of the Mandarin Chinese is the way that we build the system. For ongoing research to further improve the quality, we are trying a different prosodic modification algorithm that does not need precise pitch markers and can be smoothly interpolated. In the mean time, we are designing an assessment method to evaluate the intelligibility and naturalness of the system. This also can help us to investigate the influences of prosodic weighting and contextual weighting.

From this preliminary work, we have introduced how to integrate different technologies to significantly improve the naturalness of a speech synthesis system. The statistics and search algorithms never do well with a bad modeling. Only with more phonetic observations and knowledge, the power of modern speech technologies can be more effectively created. This is also the guideline for the development of our TTS system.

5 REFERENCES

[1] Fu-chiang Chou, Chiu-yu Tseng, Keh-jiann Chen, Lin-shan Lee, "A Chinese Text-to-Speech System Based on Part-of-Speech Analysis, Prosodic Modeling and Non-Uniform Units", *International Conference on Acoustics, Speech, and Signal Processing*, pp. 923-926,1997

[2] Shaw-hwa Hwang, Sin-horng Chen and Yih-ru Wang, "A Mandarin Text-to-Speech System", *International Conference on Spoken Language Processing*, pp. 1421-1424, 1996

[3] Ren-Hua Wang, Qinfeng Liu and Difei Tang, "A New Chinese Text-to-Speech System with High Naturalness", *International Conference on Spoken Language Processing*, pp. 1441-1444, 1996

[4] Chilin Shih and Richard Sproat, "Issues in Text-to-Speech Conversion for Mandarin", *Computational Linguistics and Chinese Language Processing vol.1, No.1* pp. 37-86,1996

[5] Sagisaka Y., Kaiki N., Iwahashi N. and Mimura. K. "ATR v-Talk Speech Synthesis System", *International Conference on Spoken Language Systems*, pp. 483-486, 1992

[6] Thomas Portele, Florian Hofer and Wolfgang J. Hess. "A Mixed Inventory Structure for German Concatenative Synthesis", *Progress in Speech Synthesis*, pp. 263-277, Springer Verlag, 1996

[7] Nick Campbell and Alan W. Black, "Prosody and the Selection of Source Units for Concatenative Synthesis", *Progress in Speech Synthesis*, pp. 279-282, Springer Verlag, 1996

[8] Alistair D. Conkie and Stephen Isard, "Optimal Coupling of Diphones", *Progress in Speech Synthesis*, pp. 294-304, Springer Verlag, 1996

[9] Chiu-yu Tseng, "A Phonetically Oriented Speech Database for Mandarin Chinese", *International Congress of Phonetic Sciences*, pp. 326-329, 1995

[10] Brugnara, D. Falavigna and M. Omologo, "Automatic Segmentation and Labeling of Speech Based on Hidden Markov Models", *Speech Communication* (12), pp. 357-370, 1993