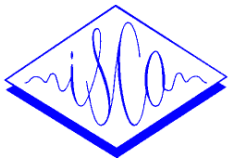

International Symposium on Chinese Spoken Language Processing (ISCSLP 2002)
Taipei, Taiwan
August 23-24, 2002

CONCATENATIVE MANDARIN TTS ACCOMMODATING ISOLATED ENGLISH WORDS

Zhenli YU, Dongjian YUE and *Jian-Cheng HUANG*

Motorola China Research Center, Shanghai
A16300@email.mot.com

ABSTRACT

An experiment to explore the method realizing a concatenative Chinese TTS accommodating isolated English words is presented. The experiment was based on an existing concatenative Mandarin TTS system, developed in Motorola China Research Center. The experimental system employs an English word synthesizer based on the concatenation of speech segments stored in an English corpus. The original English corpus contains isolated words uttered by a professional English speaker. The Chinese speaker who uttered the Mandarin TTS speech corpus uttered the same set of English words. The English word corpus uttered by the English speaker was then modified by the English word corpus uttered by the Mandarin speaker, on a word-by-word basis. A voice conversion technique is applied to modify the English word corpus. The voice conversion is focused on the voiced phones. The conversion process is basically a pitch scaling and spectral envelope scaling based on a phone level average.

1. INDTRODUCTION

Recent Chinese TTS systems tend to accommodate Chinese text mixed with isolated English words. The typical way for doing so is to allow a system alternating between two acoustically independent engines, with one serving Chinese text and the other serving English words. An obvious drawback of this approach is that the alternate voice can result in an unpleasant audio effect. In this paper, we present a new method to tackle the voice discontinuity problem based on a voice conversion technique. The proposed method combines a concatenative Mandarin TTS system with an English isolated word synthesizer. The voice of the English synthesizer was converted to the voice of the Mandarin system, allowing users to perceive a single speaker voice in both Chinese and English read-out.

Our experiment was based on the existing concatenative Mandarin TTS system, MRhyker, developed in Motorola China Research Center [1]. Accompanied with MRhyker is an English word synthesizer based on the concatenation of speech segments stored in an English corpus. Originally, the English speech corpus contained about 2,000 words (referred to as EE-corpus) uttered by a professional English speaker. The Chinese speaker who uttered the MRhyker speech corpus uttered the same set of words used in the EE-corpus, creating a corpus, called CE-corpus. The EE-corpus was then modified by the CE-corpus, on a word-by-word basis.

A voice conversion technique is applied to modify EE-corpus using CE-corpus. Considering that speaker feature is mainly conveyed in the voiced part of speech, the conversion is needed only on the voiced phones (V-phones). The conversion process is basically a pitch scaling and spectral envelope scaling based on a phone level average.

To demonstrate the conversion effect, we applied parametric speech coding to speech segments in all corpora. The coding scheme used in the experiment is the harmonic vector excitation coding (HVXC) [2]. We detected V-phones in the EE-corpus and CE-corpus, respectively, and grouped them into pairs on a word-by-word and phone-by-phone basis. The mean LSF vector, mean pitch value, and spectral mean of LP residual signal of the V-phone pairs are analyzed. The V-phones of EE-corpus are then modified by scaling the mid-term mean values so that they have identical mid-term means as their counterparts of the CE-corpus have. As a result, the modified V-phones keep the temporal dynamics of the English speaker, while they sound like uttered by the Mandarin speaker.

The experiment shows some promise that a single speaker bi-lingual TTS can be realized, even though a real single speaker bilingual corpus is not available. Particularly, because the MRhyker is an embeddable concatenative Chinese TTS system, the present method gives a promising solution to the embedded single speaker bi-lingual TTS.

2. THE EXPERIMENT FRAMEWORK

The experiment was based on the existing concatenative Mandarin TTS system, MRhyker [1], developed in Motorola China Research Center. The entire experimental flowchart consists of four major parts, English/Chinese separator, Chinese concatenative acoustic parameter arrays generator, isolated English word acoustic parameter arrays generator and concatenative waveforms synthesizer. The flowchart is shown in figure-1.

The mixed English words are detected from the input text with the English/Chinese separator module. The Chinese text is processed to select the acoustic units following the major procedures of MRhyker. This part results in the acoustic parameter arrays for concatenation for the Chinese text. The acoustic parameter arrays for concatenation for English word are also selected in parallel. At the stage of concatenation, the acoustic parameters arrays of the Chinese text and the isolated English word are concatenated and the output waveforms of the full text are then synthesized with the concatenative waveforms synthesizer.

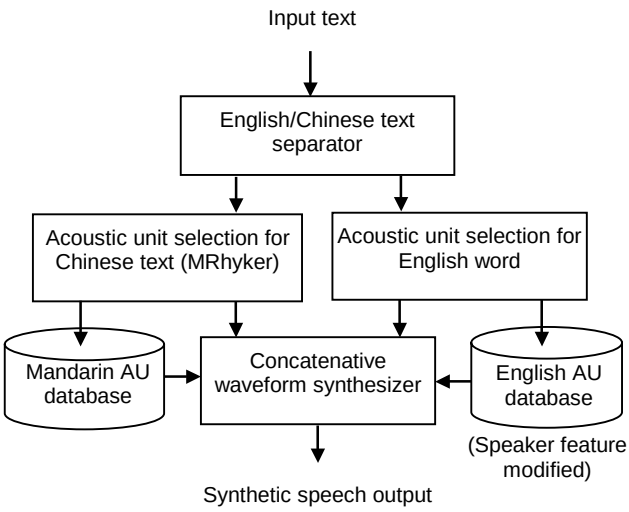

Figure-1 The experimental flowchart

3. VOICE CONVERSION OF ENGLISH WORD CORPUS

3.1 English word corpora

Originally, the English speech corpus contained about 2,000 most frequently used isolated words (referred to as EE-corpus) uttered by a professional English speaker. The Chinese speaker who uttered the MRhyker speech corpus uttered the same set of words used in the EE-corpus, creating a corpus, called CE-corpus. Both the CE-corpus and the EE-corpus are segmented at the phone level.

Although the Chinese speaker was required to utter the same set of words as the English speaker does, the English speaking quality of the Chinese speaker may not be as good as the professional English speaker. To allow users perceive the read-out of both Chinese text and English words as sounded by a single and professional speaker, voice conversion technique is applied to modify the EE-corpus using the speaker features of the CE-corpus.

3.2 Speaker feature modification

Assuming that speaker feature is mainly conveyed in the voiced portion of speech, the conversion is needed only on the voiced phones (V-phones). The conversion process is basically a pitch scaling and spectral envelope scaling based on a phone level average.

To demonstrate the conversion effect, we applied parametric speech coding to speech segments in all corpora. The coding scheme used in the experiment is the harmonic vector excitation coding (HVXC). V-phones are detected in the EE-corpus and CE-corpus, respectively, and grouped into pairs on a phone-by-phone (referred as CE-V-phone and EE-V-Phone) basis.

Spectral analysis, pitch estimation are conducted with the analysis module of HVXC. Glottal wave (residual signal) is also estimated (as shown in figure-2).

3.2.1 Pitch modification

For a V-Phone pair, both the EE-V-phone and the CE-V-phone is segmented T^e and T^c frames. The pitch trajectories of the V-phone pair are expressed as,

$$P_t^e, \quad t = 1,\cdots T^e \tag{1}$$

and

$$P_\tau^c, \quad \tau = 1,\cdots T^c \tag{2}$$

where t and τ is the frame index of the EE-V-phone and the CE-V-phone, respectively. The mid-term mean pitch value of the V-phone pairs are calculated as

$$\overline{P^c} = \frac{1}{T^c}\sum_{\tau=1}^{T^c} P_\tau^c \tag{3}$$

and

$$\overline{P^e} = \frac{1}{T^e}\sum_{t=1}^{T^e} P_t^e \tag{4}$$

The pitch value of the EE-V-phone is modified as,

$$P_t^x = \beta \cdot P_t^e, \quad t = 1,\cdots T^e \tag{5}$$

where the scaling factor for the modification is obtained from

$$\beta = \varphi \cdot \overline{P^c} \Big/ \overline{P^e} \tag{6}$$

Here, φ is a global average pitch adjustment factor. During the experiment, it was found that the global average pitch value of the CE-corpus is different from that of the Mandarin speech corpus of MRhyker. The reason is that the English speaker's global average pitch value is lower than the Mandarin speaker. When the Mandarin speaker uttered the CE-corpus, she was required to follow the English speaker's pronunciation. As a side effect, she uttered the CE-corpus with global average pitch value lower than her normal one because she tried to follow the English speaker in all aspects of pronunciation, including the pitch level.

3.2.2 Vocal-tract parameter modification

Let $X_{t,i}$ be the $i'th$ line spectral frequency (LSF) of the $t'th$ frame of a phone, the LSF vector trajectories of a V-phone pair, CE-V-phone and EE-V-phone, are represented, respectively, as,

$$R_\tau^c = \{ X_{\tau,i},\ i = 1,\cdots M \}, \quad \tau = 1,\cdots T^c \tag{7}$$

and

$$R_t^e = \{ X_{t,i},\ i = 1,\cdots M \}, \quad t = 1,\cdots T^e \tag{8}$$

The mid-term mean LSF vector of the V-phone pairs are calculated as

$$\overline{X_i^c} = \frac{1}{T^c} \sum_{\tau=1}^{T^c} X_{\tau,i},\ i = 1,\cdots M \tag{9}$$

and

$$\overline{X_i^e} = \frac{1}{T^e} \sum_{t=1}^{T^e} X_{t,i},\ i = 1,\cdots M \tag{10}$$

The vocal-tract parameters of the EE-V-phone is then modified as,

$$X_{t,i}^m = \alpha_i \cdot X_{t,i}^e,\ i = 1,\cdots M \tag{11}$$

where α_i is the scaling factor for the vocal-tract parameters modification that is obtained from

$$\alpha_i = \overline{X_i^c} \Big/ \overline{X_i^e},\ i = 1,\cdots M \tag{12}$$

3.2.3 Glottal wave modification

From the original residual signal, g_t^e, of the EE-V-phone, a pitch-modified residual signal, g_t^x, which pertains the modified pitch trajectories P_t^x, $t = 1,\cdots T^e$, is re-synthesized using HVXC. The spectral trajectories of the residual signal of the CE-V-phone, called $Y_{\tau,i}^c$, and the spectral trajectories of the pitch-modified residual signal of the EE-V-phone, called $Y_{t,i}^x$, are obtained from g_τ^c and g_t^x. The spectral trajectories of the pitch-modified residual signal is then modified as,

$$Y_{t,i}^m = \eta_i \cdot Y_{t,i}^x,\ i = 1,\cdots M;\ t = 1,\cdots T^e \tag{13}$$

where the scaling factor for the spectral mean modification is obtained from

$$\eta_i = \overline{Y_i^c} \Big/ \overline{Y_i^x} \tag{14}$$

After then, a final residual signal g_t^m, in which both pitch and spectral mean are scaled, is then obtained by doing the inverse Fourier transform of the spectral trajectories of $Y_{t,i}^m$.

With the LSP-modified vocal-tract parameters $X_{t,i}^m$ and the pitch-spectral-modified residual signal g_t^m, a new speech signal of the EE-V-phone is re-synthesized. In this new EE-V-phone signal, the speaker features, mainly characterized by mean pitch value and mean vocal-tract parameters, are modified so that it has identical mean values as its counterparts of the CE-V-phone has. As a result, the new EE-V-phone keeps the temporal dynamics of the English speaker, while it sounds like uttered by the Mandarin speaker.

The entire procedure of the voice conversion of English word corpus is shown in figure-2.

By repeating the above procedure of modification for all the V-phones of the EE-corpus, the voice conversion of the EE-corpus is realized. As a final result, the modified words of EE-corpus keep the temporal dynamics of the English speaker, while they sound like uttered by the Mandarin speaker. This new EE-corpus, called M-EE-corpus, is then used as the acoustic unit database for the English word synthesizer.

4. WORD CONCATENATION

To use a small speech inventory to construct arbitrary English words, we apply unit concatenation in our experiment.

To do this, an acoustic unit database is produced from the acoustic segmentations of the M-EE-corpus. If an isolated word to be read-out is found in the word list of the corpus, it is constructed by concatenating the corresponding acoustic units of the word from the database. Otherwise, the word is constructed by concatenating the best acoustic unit candidates selected from the unit database. The selection of the best unit candidates is based on phonetic property analysis and optimal search algorithm.

The detailed description of word construction based on unit concatenation, however, is not the topic discussed in this paper and will not be elaborated here.

5. DEMO SAMPLE

To demonstrate the synthesis result of the concatenative Chinese TTS accommodating isolated English words, the following text is selected for benchmarking.

“有人将超市叫做 Supermarket，将笔记本电脑称做 Notebook，把教授尊称为 Professor。有些大学毕业生把录用通知称做 Offer，将办公室叫做 Office。而一些缩略语也是大行其道，如把博士学位说成 Ph.D.，至于 WTO 则连普通老百姓也能脱口而出。” (Example)

6. CONCLUSIONS

An experiment to explore the method realizing a concatenative Chinese TTS accommodating isolated English words is presented.

The experiment was based on an existing concatenative Mandarin TTS system. The experimental system employs an English word synthesizer based on the concatenation of speech segments stored in an English corpus. The original English

corpus contains isolated words uttered by a professional English speaker. The Chinese speaker who uttered the Mandarin TTS speech corpus uttered the same set of words. The English word corpus uttered by the English speaker was then modified by the English word corpus uttered by the Mandarin speaker, on a word-by-word basis.

A voice conversion technique is applied to modify the English word corpus. The voice conversion is focused on the voiced phones. The conversion process is basically a pitch scaling and spectral envelope scaling based on a phone level average.

The experiment shows promise that a single speaker, bi-lingual, concatenative TTS can be realized using a small speech inventory, even though a real single speaker bi-lingual corpus is not available.

7. REFERENCES

[1] Fang Chen, Gui-Lin Chen, Jian-Cheng Huang, Zhen-Li Yu, Dong-Jian Yue, Yi-Qing Zu, "Natural Sounding Embedded Text-To-Speech Systems," Proc. of the 5th National Conference on Modern Phonetics, Beijing, China, Oct. 2001

[2] Dongjian Yue, Jian-cheng Huang and Zhenli Yu, "An Experiment of a High Quality Speech Synthesizer Based on Harmonic Vector Excitation Coding," Proc. of NCMMSC'2001, Shenzhen, China, Nov. 2001

[3] E. Moulines and Y. Sagisaka (edited), *Voice conversion: State of the art and perspectives, Special issue of Speech communication*, 16(2), 1995

[4] Kleijn, W.B. and Paliwal, K.K. (edited), *Speech coding and synthesis*, The Netherlands: Elseevier, 1995

[5] Flanagan, J.L., *Speech analysis, synthesis, and perception*, 2nd ed., New York: Springer-Verlag, 1972

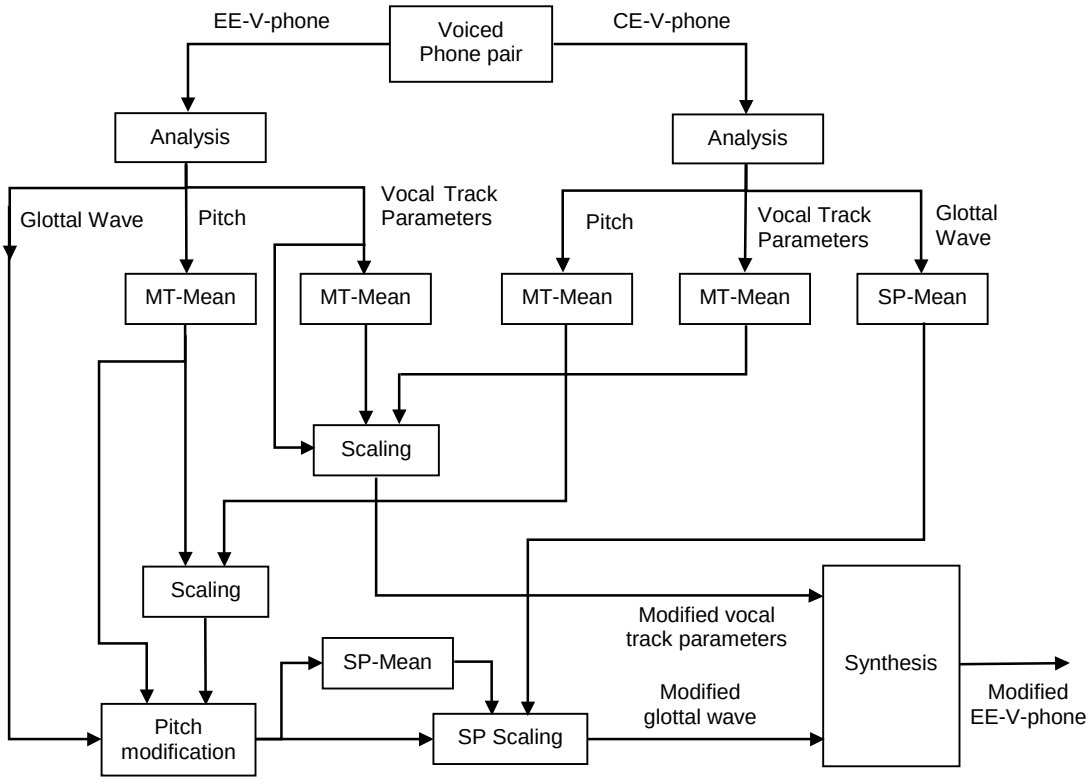

Figure-2 Procedure of voice conversion for English word corpus