ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

AN ACOUSTIC AND ARTICULATORY KNOWLEDGE INTEGRATED METHOD FOR IMPROVING SYNTHETIC MANDARIN SPEECH'S FLUENCY

Hung-Yan Gu* and Kuo-Hsian Wang#

*Department of Computer Science and Information Engineering, #Department of Electrical Engineering
National Taiwan University of Science and Technology, Taipei
e-mail: guhy@mail.ntust.edu.tw http://www.csie.ntust.edu.tw

ABSTRACT

In synthetic Mandarin speech, discontinuity of formant traces at syllable boundaries is a key factor that lowers fluency level. Therefore, we study an acoustic and articulatory knowledge integrated method to solve this discontinuity problem. First, representative trisyllable contexts are selected and their signals are recorded. The middle syllable's signal of each trisyllable pronunciation is then extracted to make a synthesis unit. To select a synthesis unit among multiple candidates, a distance function is defined to measure the spectral similarity between two synthesis units to be concatenated. In addition, several linking-restriction rules are derived, according to articulatory knowledge, to prevent some synthesis units being linked into a sequence. Then, a globally best synthesis-unit sequence is searched by using a dynamic programming based algorithm. When the method above is applied, the formant traces at syllable boundaries will become smoother. Also, subject evaluation shows that the fluency level of synthetic Mandarin speech can indeed be improved a lot.

1. INTRODUCTION

The quality of a synthetic speech is usually evaluated in such issues as naturalness, intelligibility, and fluency. Naturalness evaluates whether a synthetic speech is as natural as spoken by a person and has no machine accent. Intelligibility evaluates how many percents of words in synthetic speeches can be caught. And fluency evaluates whether a synthetic speech is as fluent as spoken by a person. Fewer prosodic and acoustic discontinuities lead to higher fluency level. Many researchers (including us) had studied and constructed different prosodic models in the past [1, 2, 3, 4, 5, 6, 7]. They intend to improve the quality of synthetic speech from the processing of prosodic parameter generation. These efforts indeed make naturalness and intelligibility highly promoted. However, improvements in fluency are more diverse among researchers.

The issue, fluency, can be subdivided into prosodic and acoustic fluencies [8]. Prosodic fluency is closely related to naturalness. It evaluates continuity of prosodic characteristics within and between syllables. For examples, sudden change of pitch-contour height or intensity between two adjacent syllables will be perceived as prosodic discontinuity. On the other hand, acoustic fluency evaluates continuity of acoustic characteristics. For example, disconnected formant-traces at syllable boundary will be perceived as acoustic discontinuity. Although a good prosodic model can offer prosodic fluency well, it however does not help for acoustic fluency. Take our previous version of Mandarin text-to-speech system as an example, which can be on-line tested from the web site, **http://guhy.ee.ntust.edu.tw/gutts/**. Its prosodic model can offer a certain level of prosodic fluency (in pitch contour especially). However, it does not solve the acoustic-fluency issue of smoothening formant-trace transition.

Take the short sentence, /tai-2 uan-1 ke-1 zi-4/ ("台灣科技"), as an example. If it is synthesized by our previous system and analyzed, the spectrum would be as shown in Fig. 1. If an utterance recorded from a person is analyzed instead, the spectrogram would be as shown in Fig. 2. In both Fig. 1 and 2, we can see deep colored formant traces, named F1, F2, F3, etc., in order [8]. Compare these two figures, we can find that the F2 trace of /tai-2/ goes down to approach the F2 trace of /uan-1/ in Fig. 2 whereas the F2 trace of /tai-2/ goes up and apart from the F2 trace of /uan-1/ in Fig. 1. In addition, the F2 traces of /ke-1/ and /zi-4/ approach each other in Fig. 2 whereas the corresponding F2 traces in Fig. 1 goes horizontally parallel. That is, the formant traces at the boundary of /tai-2/ and /uan-1/ and at the boundary of /ke-1/ and /zi-4/ are discontinuous in Fig. 1 but are smoothly transited in Fig. 2. Therefore, we think formant trace discontinuity is a key factor that accounts for the worse acoustic-fluency presented in synthetic speech.

In our previous system, syllable unit is adopted as synthesis unit and each syllable is recorded only once. Suppose that acoustic-fluency improving is to be accomplished with only one recording of each syllable. Then, we will need an articulation model to plan a smoothed transition path for the formant traces across the syllable boundaries. Also, we need a signal processing method to shift a formant traces to theirs planed paths. However, such articulation model and formant-trace shifting method are not readily developed and not available now. Therefore, we consider recording a same syllable several times under different contexts, to have several candidate synthesis units for each syllable. Then, for a syllable to be synthesized, a synthesis unit extracted from a context that is most like to the target context can be selected. In this way, we think smoother transitions of formant traces around syllable boundary can be obtained. In this paper, we define a context of a syllable, Y, as constructed from three consecutively uttered syllables, X, Y, Z. That is, the concerned syllable Y is front and back appended with two independent syllables X and Z. X is called the front-connecting syllable and Z is called the back-connecting syllable.

The approach, developing good synthesis unit selecting algorithm to achieve higher level of fluency and naturalness, has been adopted by many researchers [9, 10, 11, 12]. Here, we also adopt this approach but only focus on improving acoustic fluency

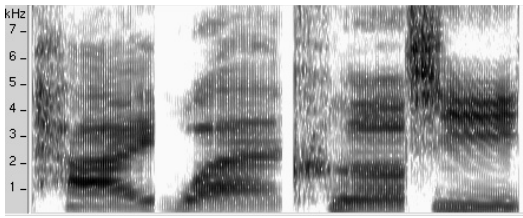

Fig. 1 Spectrogram for a synthetic speech of "台灣科技".

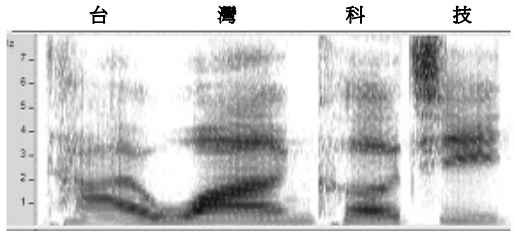

Fig. 2 Spectrogram for an uttered speech of "台灣科技".

at syllable boundaries. As to prosodic fluency, we trust to our previously studied HMM based model [6], or we can trust to a good prosodic model developed by another researcher. We think it is more practical (to implement) and more expectable (to work) that prosodic and acoustic fluencies are considered and solved separately. Apparently, when the two kinds of fluencies are to be solved simultaneously (e.g. efforts in the works [9, 13]), tremendous number of contexts and their utterances need to be prepared in order to treat the tremendous possible combinations of prosodic and acoustic factors (i.e. a corpus based approach). If only limited quantity of contexts and utterances are prepared, some kinds of compromise processing will be performed, which will inevitably decrease the fluency and quality of the synthesized speech. Therefore, the approach that solves prosodic and acoustic problem separately is meaningful.

2. SYLLABLE CONTEXT

In this paper, we assume the pitch-contour of a syllable uttered in the first tone can be modulated to obtain another tone's pitch-contour by the waveform synthesizing unit. Hence, the lexical tones need not be distinguished when counting the possible contexts of a syllable. Since there are 409 different syllables in Mandarin, each syllable may have as many as 409 × 409 = 167,281 contexts. But for implementation consideration, we have to reduce the number of contexts and to keep only the representative contexts for a syllable.

Consider first how to select front-connecting and back-connecting syllables, X and Z, in order to form contexts, XYZ. After check and compare the formant frequency values of F1 and F2 for all vowels, we decide to place the three simple syllables, /a, i, u/, as candidate for X and Z, respectively. This is because these three vowels are respectively located at the three corners of the vowel triangle in F1-F2 plot [8]. A vowel if locating at a corner of the triangle would have maximum or minimum values for F1 and F2. And such values would cause a large and representative transition of formant traces at the boundary with the middle syllable Y. Transition of formant traces caused by syllable-final of nasal ending is also considered to be typical in Mandarin. Hence, we select the syllable, /an/, as a candidate for the front-connecting syllable X. On the other hand for the back-connecting syllable Z, we select the syllable /ma/ as the representative for those syllables of an initial nasal. Also, we select the two syllables, /sa, ba/, as the representatives for those syllables of an initial consonant that is either long or short in duration. Accordingly, we have 4 candidates, /a, i, u, an/, for the front-connecting syllable X and 6 candidates, /a, i, u, ma, sa, ba/, for the back-connecting syllable Z. Therefore, we need to record 9,816 (4 × 409 × 6) trisyllable contexts' pronunciations. The sampling rate is 22,050Hz and the resolution is 16bits/sample.

For each context's pronunciation, we need to determine and label the syllable-boundary points before and after the middle syllable. According to the labeled points, the signal waveform of the middle-syllable can then be extracted out for waveform synthesizing. In this paper, for a context XYZ, the boundary points between X and Y and between Y and Z are determined in two stages. First, the boundary points are automatically selected by a segmentation program written by us. Then, the selected points are manually checked and corrected. From the results of preliminary experiments, the quality (signal clarity and acoustic fluency) of the synthesized speech will be significantly degraded if the boundary points are determined incorrectly. Also, our program cannot prevent selecting wrong boundary points. Hence, we must manually check and correct the boundary points selected by program.

3. SYNTHESIS UNIT SELECTION

In this paper, synthesis unit selection is performed for each sentence to be synthesized as a whole, and by using a DP (dynamic programming) based algorithm to select a globally best sequence of synthesis units. In the DP algorithm, we define an acoustic spectral distance function to measure the smooth-transition level of adjacent synthesis units' formant traces. In addition, we derive some linking-restriction rules according to articulatory knowledge of phonemes. Some adjacent synthesis units cannot be linked because there will be articulatory discontinuity at their boundary, which may not be faithfully detected with spectral distance function.

3.1 Spectral Distance Function

Some distance functions to measure spectral similarity between two synthesis units had been proposed in previous studies by others [10, 11]. Here, we base on those studies to design a distance function to meet our need. In details, suppose s_i and s_{i+1} are candidate synthesis units for the i'th and (i+1)'th syllable respectively. The spectral distance between s_i and s_{i+1} is measured as

$$C(s_i, s_{i+1}) = w_b \bullet D_b(s_i, s_{i+1}) + w_c \bullet D_c(s_i, s_{i+1}) \quad (1)$$

In this equation, $D_b(s_i, s_{i+1})$ is the distance component from matching the boundary frames between s_i and s_{i+1}, and $D_c(s_i, s_{i+1})$ is the distance component from matching the middle frames of s_i and s_{i+1}. w_b and w_c are weighting factors, which are set to 2 and 1 respectively. The definitions for $D_b(s_i, s_{i+1})$ and $D_c(s_i, s_{i+1})$ are

$$D_b = \sum_{1 \le n \le 3} d(f_{n+6}, g_n) \quad (2) \qquad D_c = \sum_{4 \le n \le 6} d(f_n, g_n) \quad (3)$$

In these equations, f_n represents the n'th frame of s_i, g_k represents the k'th frame of s_{i+1}, and $d(f_n, g_k)$ means a geometric distance measure on the MFCC feature vectors of f_n and g_k. Note the

range of the index variable, n, in f_n is 1 to 9. The values 1 to 3 are meant that the frames f_n are taken from the left boundary of s_i, the values 4 to 6 are meant the frames are taken around the syllable middle point, and the values 7 to 9 are meant the frames are taken from the right boundary.

3.2 Best Sequence Selection with DP

Suppose the sentence to be synthesized comprises N syllables. Then, the number of possible synthesis-unit sequences that can be linked out is 24^N. Therefore, we cannot search for the best sequence in an exhaustive manner but must resort to a DP based algorithm. Let us define $E(i, j)$ as the minimum accumulated distance of all sequences that stop at the j'th candidate synthesis unit of the i'th syllable and come from different synthesis unit for the syllables with indices less than i. Then, the recurrence relation below can be obtained according to the definition of $E(i, j)$.

$$E(i,j) = \min_{1\le k\le 24}\left[E(i-1,k) + C(s_{i-1,k}, s_{i,j})\right] \quad (4)$$

Here, $s_{i,j}$ represent the j'th candidate synthesis unit of the i'th syllable, and the definition of $C(x,y)$ is as in Equation (1). With Equation (4), the accumulated distance of a best synthesis-unit sequence can be computed as

$$E_{min} = \min_{1\le k\le 24}\left[E(N,k)\right] \quad (5)$$

Also, the best synthesis unit for the last syllable can be determined as the one with the index as

$$K = \mathbf{argmin}_{1\le k\le 24}\left[E(N,k)\right] \quad (6)$$

To back track the sequence of synthesis units that accumulates to a minimum distance as computed in equation (5), we must designate a variable, e.g. $R(i,j)$ here, to memorize which value of the index k in Equation (4) accumulates to a minimum value as saved in $E(i,j)$. Thus, the value of $R(i,j)$ is defined as

$$R(i,j) = \mathbf{argmin}_{1\le k\le 24}\left[E(i-1,k) + C(s_{i-1,k}, s_{i,j})\right] \quad (7)$$

In terms of the link information saved in $R(i,j)$, we can then back track from the K synthesis unit of the last syllable to find the best sequence of synthesis units.

3.3 Linking-Restriction Rules

If only spectral distances are used in the DP best sequence searching algorithm, the synthesized speech will still be perceived with abrupt changes (influent) at syllable boundaries sometimes. This is thought to be due to discontinuous articulator motion around syllable boundaries. Therefore, we decide to integrate articulatory knowledge of phonemes into the DP algorithm to guide the searching process. In practice, we set up linking restriction rules to prevent two synthesis units coming from contradictory contexts being linked into a sequence.

3.3.1 Sentence-Start Rule

The synthesized speech will be perceived as not pronounced from silence and having strange articulator motion if a synthesis unit of coarticulation effect in its leading phonemes is used to synthesize the starting syllable of a sentence. Note that among the four front-connecting syllables, /a, i, u, an/, the first three are single vowels and may easily be coarticulated with their following syllable in a context. Thus, we set up the rule:

(Rule 1): *A synthesis unit, **yy**, to be used for the starting syllable must be extracted from a context of the form, **an_yy_zz**. Here, **zz** may be any of the 6 possible back-connecting syllables.*

3.3.2 Sentence-End Rule

If a synthesis unit of coarticulation effect in its last phonemes is used to synthesize the ending syllable of a sentence, the synthesized speech will be perceived as not terminated to silence at the end, and may be perceived as suddenly cut. Note that among the six back-connecting syllables, /ba, sa, ma, a, i, u/, only the syllable, /ba/, is started with a mouth-closed phoneme. Thus, we set up the rule:

(Rule 2): *A synthesis unit, **yy**, to be used for the ending syllable must be extracted from a context of the form, **xx_yy_ba**. Here, **xx** may be any of the 4 possible front-connecting syllables.*

3.3.3 Motion-Continuity Rule

First, the sizes of mouth opening are divided into three classes, large-mouth, medium-mouth, and small-mouth. A syllable is defined as started or ended with large-mouth if it is started or ended with one of the phonemes /a, o/. Another case if a syllable is started or ended with the phoneme /i/ or any one consonant, it is defined as started or ended with small-mouth. For the other cases, a syllable is defined as started or ended with medium-mouth. In terms of the mouth-opening sizes defined above, we derive motion-discontinuity checking rules to prevent articulatory discontinuities. Suppose the two synthesis units s_{i-1} and s_i are to be linked. We check if the back-connecting syllable in the context of s_{i-1} is started with large-mouth (small-mouth) while the front-connecting syllable in the context of s_i is ended with small-mouth (large-mouth). If any of the conditions checked does occur, motion discontinuity is then detected. Apparently, changing mouth's size abruptly from large into small is discontinuous motion and will not occur in practice.

A more practical example is as shown in Fig. 3, which shows possible links between synthesis units for the short sentence, /tai uan/. Due to the sentence-start rule, the candidate synthesis units for /tai/ must be selected from contexts that include /an/ as front-connecting syllable. Similarly, due to the sentence-end rule, the candidate synthesis units for /uan/ must be selected from contexts that include /ba/ as back-connecting syllable. In addition, consider the synthesis unit for /uan/ that comes from the context /a_uan_ba/. Since this unit's front-connecting syllable /a/ is ended with large-mouth, this unit cannot be linked to a former synthesis unit that has a small-mouth started back-connecting syllable such as /tai/ from /an_tai_i/ or /an_tai_ba/.

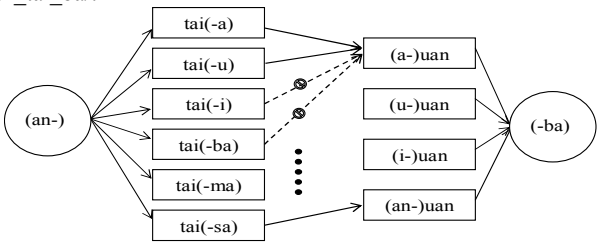

Fig. 3 Possible links between the synthesis units for synthesizing /tai uan/

In practical implementation, the restriction of the rules in Section 3.3 when detected can be converted to a very large

distance and added to the distance, $C(s_i, s_{i+1})$, computed with Equation (1). Then, the DP based best sequence searching algorithm discussed in Section 3.2 can still be applied. But now articulatory knowledge is utilized in addition to acoustic knowledge.

4. EXPERIMENTAL EVALUATION

To evaluate the method studied here, we have to synthesize some example speech for listening tests. In the text-to-speech system used, the method TIPW [14] is used to synthesize signal waveform and the method SPC-HMM [6] is used to generate pitch contours while the other prosodic parameters are generated with rules. Subjective evaluation of the synthesized speech is made. 20 persons are invited to participate in the listening tests. For intelligibility evaluation, two short articles are synthesized and played to the listeners. Here, intelligibility level is defined as the ratio of correctly written out characters. For fluency level evaluation, another article is synthesized and played to the listeners. Each listener gives a fluency score for the synthesized speech. The score ranges for reference had been explained to the listeners beforehand. In details, (a) 90~99 means very fluent as spoken by a person; (b) 80~89 means the fluency level is good; (c) 70~79 means ordinary and acceptable; (d) 60~69 means below ordinary and is not acceptable; (e) 50~59 means very poor.

To have a reference for comparison of fluency level, a single-unit mode is also set up to synthesize speech for listening tests. In single-unit mode, the synthesis unit for a syllable, yy, is fixedly extracted from its trisyllable context of the form, /an_yy_ba/. On the other mode, i.e. multi-unit mode, the method as explained in Section 3 is used. After the articles are synthesized at both modes, the synthesized speech are then randomly permuted and played to the listeners. After computing the average rates of correctly written out Chinese characters, we obtain the rates for intelligibility test as shown in the first row of Table 1. Also, averaging the fluency scores collected, we obtained the mean scores for both modes as shown in the second row of Table 1. In intelligibility, both modes can achieve rates higher than 99%. Therefore, acoustic discontinuity at syllable boundary will not cause significant difference for intelligibility level. But for fluency, the score, 89, obtained by the multi-unit mode is apparently higher than the score, 84, obtained by the single-unit mode. Therefore, the method studied in this paper can indeed improve the fluency level of the synthesized Mandarin speech.

Table 1 Evaluation Results from Listening Tests

	Single-unit Mode	Multi-unit Mode
Intelligibility	**99.7 %**	**100 %**
Fluency	**84**	**89**

5. CONCLUSION

The discontinuities of formant trace transitions at syllable boundaries are thought to be a key factor that results in lowered acoustic fluency for synthetic Mandarin speech. Therefore, we studied and developed a method to make the formant trace transition smoother at syllable boundaries. This method is designed to integrate acoustic knowledge (spectral distance measure) and articulatory knowledge (linking restriction rules derived from phoneme-articulation knowledge) into a DP based algorithm. Then, a global best sequence of synthesis units can be found in a time-efficient way. According to the listening tests performed, the method proposed can indeed significantly improve the fluency level of the synthesized Mandarin speech. In more detailed, the fluency score of out method is 89 which is 5 points more than 84 obtained from using a single fixed synthesis unit for each syllable. We think the gap in fluency scores can be further enlarged. Note that only about half of the 9,816 trisyllable contexts are manually checked for syllable-boundary segmenting errors, and segmenting errors are frequently found and are influential for the fluency of the synthesized speech.

REFERENCE

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

[2] Lee, L. S., C. Y. Tseng and C. J. Hsieh, "Improved Tone Concatenation Rules in a Formant-based Chinese Text-to-Speech System", IEEE trans. Speech and Audio Processing, Vol. 1, pp. 287-294, 1993.

[3] Chen, S. H., S. H. Hwang, and Y. R. Wang, "An RNN-Based Prosodic Information Synthesizer for Mandarin Text-to- Speech", in IEEE trans. Speech and Audio Processing, Vol. 6, No. 3, pp. 226-239, 1998.

[4] Wu, C. H. and J. H. Chen, "Automatic Generation of Synthesis Units and Prosodic Information for Chinese Concatenative Synthesis", Speech Communication, Vol. 35, pp. 219-237, 2001.

[5] Yu, M. S., N. H. Pan, and M. J. Wu, "A Statistical Model with Hierarchical Structure for Predicting Prosody in a Mandarin Text-to-Speech System", International Symposium on Chinese Spoken Language Processing (ISCSLP 2002), Taipei, pp. 21-24, 2002.

[6] Gu, H. Y. and C. C. Yang, "A Sentence-Pitch-Contour Generation Method Using VQ/HMM for Mandarin Text-to-speech", International Symposium on Chinese Spoken Language Processing (ISCSLP2000), Beijing, pp. 125-128, 2000.

[7] Lin, C. T., R. C. Wu, J. Y. Chang, and S. F. Liang, "A Novel Prosodic-Information Synthesizer Based on Recurrent Fuzzy Neural Network for the Chinese TTS System", IEEE trans. Systems, Man, and Cybernetics, Vol. 34, No. 1, pp. 309-324, Feb. 2004.

[8] O`Shaughnessy, D., Speech Communication: Human and Machine, 2'nd ed., IEEE Press, 2000.

[9] Chou, Fu-chiang, Corpus-based Technologies for Chinese text-to-Speech Synthesis, Ph. D. Dissertation, Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, 1999.

[10] Chen, Jau-Hung, A Study on Synthesis Unit Selection and Prosodic Information Generation in a Chinese Text-to-Speech System, Ph.D. Dissertation, Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan, 1998.

[11] Toda, T., H. Kawai, M. Tsuzaki, and K. Shikano, "Unit Selection Algorithm for Japanese Speech Synthesis Based on Both Phoneme Unit and Diphone Unit", IEEE ICASSP, Orlando, USA, Vol. 1, pp. 465-468, 2002.

[12] Chu, M., H. Peng, H. Y. Yang, and E. Chang, "Selecting non-uniform units from a very large corpus for concatenative speech synthesizer", IEEE ICASSP, UT, USA,Vol. 2, pp. 785-788, 2001.

[13] Sagisaka, Y., *et al.*, "ATR v-talk Speech Synthesis System", Proc. ICSLP'92, Canada, pp. 483-486, 1992.

[14] Gu, H. Y. and W. L. Shiu, "A Mandarin-syllable Signal Synthesis Method with Increases Flexibility in Duration, Ton and Timbre Control", Proc. Natl. Sci. Counc. ROC(A), vol. 22, No.3, pp. 385-395, 1998.

ACKNOWLEDGEMENT

This study is supported by National Science Council under the contract number, NSC 90-2213- E-011-048.