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

Hung-Yan Gu and Kuo-Hsian Wang*

Department of Computer Science and Information Engineering
*Institute of Electrical Engineering
National Taiwan University of Science and Technology
e-mail: guhy@mail.ntust.edu.tw

ABSTRACT

In synthetic Mandarin speech, an important factor that lowers the fluency level is formant-trajectory discontinuities between adjacent syllables. Therefore, we propose a method that integrates acoustic and articulatory knowledge to solve this discontinuity problem. First, representative trisyllable contexts are selected and their speech signals are recorded. Then, the middle syllable's signal of each trisyllable utterance is extracted to form a synthesis unit. To select a synthesis unit among multiple candidates, we define an acoustic distance function to measure the spectral similarity between the two synthesis units to be concatenated. In addition, we derive several constrained-selection rules based on articulatory knowledge to prevent some synthesis units from being connected into a sequence. Then, a globally best synthesis-unit sequence is searched using a dynamic programming based algorithm proposed here. When the proposed method is applied, the formant trajectories at syllable boundaries will become smoother. Also, listening tests show that the fluency level of synthetic Mandarin speech can indeed be improved.

Keywords: speech synthesis, fluency level, unit selection, articulatory knowledge.

I. INTRODUCTION

A synthetic speech set's quality is usually evaluated in the three issues, *i.e.* naturalness, intelligibility, and fluency. Naturalness evaluates whether the synthetic speech is as natural as that spoken by a person and whether it has a machine accent. Intelligibility evaluates the percentage of words in a synthetic speech set that can be understood. Finally, fluency evaluates whether a synthetic speech set is as fluent as that spoken by a person. Many researchers have studied and have constructed different kinds of prosodic models in the past [1-8]. They intend to improve the quality of synthetic speech from the generation of prosodic parameters. These efforts indeed have improved the naturalness and intelligibility considerably. Improvements in fluency, however, are more diverse among different researchers' synthesis systems.

The issue, fluency, can be subdivided into prosodic and acoustic fluencies [9]. Hence, fewer prosodic and acoustic discontinuities will lead to a higher fluency level. Prosodic fluency is closely related to naturalness. It evaluates the continuity of prosodic characteristics within and between syllables. For example, a sudden change of pitch heights or intensities of two adjacent syllables will be perceived as prosodic discontinuity. On the other hand, acoustic fluency evaluates the continuity of acoustic characteristics. For example, the disconnected formant-trajectories at the syllable boundary of two adjacent syllables will be perceived as acoustic discontinuity. Although a good prosodic model can offer good prosodic fluency, it does not help for acoustic fluency. As an example, take the previous version of our Mandarin speech synthesis system, which can be tested on-line at http://guhy.ee.ntust.edu.tw/gutts/. Its prosodic model can offer a certain level of prosodic fluency (especially in pitch contour). Nevertheless,

it does not solve the issue of acoustic fluency, *i.e.* formant-trajectory transitions between adjacent syllables are not smoothed.

Take the short sentence, /tai-2 wan-1 ke-1 zi-4/ ("台灣科技"), as an example. If it is synthesized by our previous system and the synthetic speech signal is analyzed, the spectrogram obtained would be the one shown in Fig. 1. Instead, if an utterance of the same sentence from a person is recorded and analyzed, the spectrogram obtained would be the one shown in Fig. 2. In both Figure 1 and Figure 2, we can see deep colored formant trajectories that are conventionally named F1, F2, F3, *etc*. [9]. Comparing these two figures, we find that the F2 trajectory of /tai-2/ goes down to approach the F2 trajectory of /wan-1/ in Fig. 2, whereas the F2 trajectory of /tai-2/ goes up and apart from the F2 trajectory of /wan-1/ in Fig. 1. In addition, the F2 trajectories of /ke-1/ and /zi-4/ approach each other in Fig. 2, whereas the corresponding F2 trajectories in Fig. 1 go horizontally parallel. That is, the formant trajectories at the boundary between /tai-2/ and /wan-1/ and at the boundary between /ke-1/ and /zi-4/ are discontinuous in Fig. 1 but are smoothly transited in Fig. 2. Therefore, we think formant trajectory discontinuity is an important factor that accounts for the lower acoustic fluency presented in a synthetic speech.

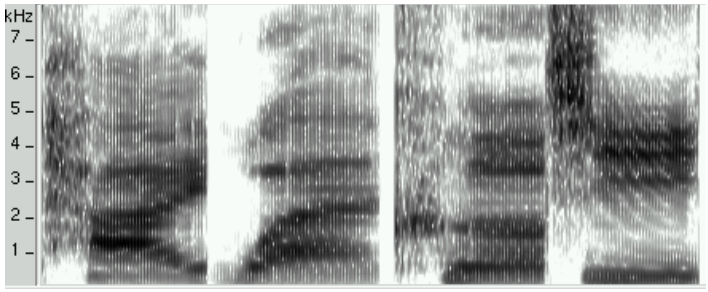

Fig. 1. Spectrogram analyzed from a previously synthesized speech.

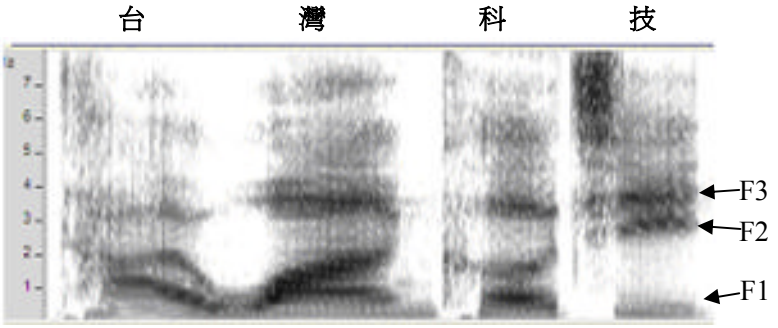

Fig. 2. Spectrogram analyzed from a person's uttered speech.

In our previous system, the syllable is adopted as the synthesis unit, and each Mandarin syllable is recorded only once. Suppose that acoustic fluency is to be improved with just one recorded utterance for each Mandarin syllable. Then, we would need an articulation model to plan a smoothed transition path for each formant trajectory around the boundary of two adjacent syllables. Also, we would need a signal processing method to adjust formant trajectories. Nevertheless, such an articulation model and formant-trajectory adjusting method are either under development or have not been verified practically by a speech synthesis system [10]. Therefore, we consider another approach that records the same syllable several times under different contexts to have several candidate synthesis units for each Mandarin syllable. Then, when a syllable is to be synthesized, a synthesis unit extracted from the context that is most like the target context can be selected. In this manner, we think smoother transitions of formant trajectories

around a syllable boundary can be obtained. In this paper, we define a context as a sequence of three independent syllables, X, Y, and Z, that are artificially connected to form a trisyllable, XYZ. The trisyllable will be uttered continuously without pauses placed in between. That is, the concerned syllable Y can be preceded and succeeded with two independent syllables, X and Z, to form a context for Y. Hence, X is the context-preceding syllable and Z is the context-succeeding syllable.

The approach, developing a good synthesis-unit selecting algorithm to achieve a higher level of fluency and naturalness, has been adopted by many researchers [11-15]. Here, we also adopt this approach but only focus on improving acoustic fluency at syllable boundaries. As to prosodic fluency, we trust our previously studied HMM based model [6], or we can trust a good prosodic model developed by another researcher. In addition, consider that the functions of variable speaking rate and voice-timbre transformation (*e.g.* transforming a female adult's timbre into a male adult's timbre) are intended to be provided. Therefore, we think that improving prosodic and acoustic fluencies separately is more practical to implement and more expectable to work. When the two kinds of fluencies are to be solved simultaneously (*e.g.* efforts in [11, 14, 16]), a tremendous number of contexts and their utterances need to be prepared and recorded in order to treat the tremendous number of possible combinations of prosodic and acoustic factor values (including different speaking rates). If only a limited quantity of contexts and utterances are prepared and recorded, some compromises must be made, which will inevitably decrease the fluency and quality of the synthetic speech. This problem apparently becomes more serious when the functions of variable speaking rate and voice-timbre transformation need also be considered. Therefore, an approach that solves prosodic and acoustic problems separately is thought to be more practical. Recently, we have applied this approach to synthesize Mandarin speech. In the future, we consider applying the technique developed here to synthesize speech of other languages, *e.g.* Min-nan and Hakka [17].

II. 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. Hence, the five lexical tones of Mandarin 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 $409 \times 409 = 167{,}281$ contexts. Nevertheless, for implementation consideration, we have to reduce the number of contexts and keep only the representative contexts for a syllable.

Consider first how to select context preceding and succeeding syllables, X and Z, in order to form a trisyllable, XYZ, for a concerned syllable Y. After checking and comparing the formant frequency values of F1 and F2 for all vowels, we decided to place the three simple syllables, /a, i, u/, as candidates for X and Z, respectively. This is because these are the three vowels located at the three corners of the vowel triangle in the F1-F2 plot [9]. A vowel located at a corner of the triangle would have maximum or minimum frequency values for F1 and F2. Such F1 and F2 values would cause a large and representative transition of formant trajectory at the boundaries between X and Y and between Y and Z. In addition, transitions of formant trajectories caused by syllable-final nasal endings are also considered to be typical in Mandarin. Hence, we select the syllable, /an/, as a candidate for the context-preceding syllable X. On the other hand, for the context-succeeding 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 context-preceding syllable X and 6 candidates, /a, i, u, ma, sa, ba/, for the context-succeeding syllable Z. Therefore, we need to record 9,816 ($4 \times 409 \times 6$) trisyllable contexts' utterances. Here, the sampling rate is set to 22,050Hz and the resolution is 16bits/sample.

In each context's recorded speech signal, we have to label the left and right boundary points for the middle syllable. According to the labeled points, the signal of the middle-syllable can then be extracted for signal synthesis processing. In this paper, the boundary points between X and Y and between Y and Z for a context XYZ are labeled in two steps. First, the boundary points are automatically selected by a segmentation program developed in our laboratory. Then, the selected points are manually checked and corrected. According to the results of the preliminary experiments, the quality (signal clarity and acoustic fluency) of the synthesized speech will be significantly degraded if the boundary points are labeled incorrectly. Also, our program cannot avoid selecting wrong boundary points. Hence, we must manually check and correct the boundary points selected by the segmentation program.

In implementing the segmentation program, we adopted an idea proposed in Huang's thesis [18], which is explained in the following. When given a trisyllable signal file, SIG_X_Y_Z, recorded for a context, XYZ, we can know its syllable content from its file name. Therefore, we first concatenate the three signal files that are recorded in isolation for the three syllables, X, Y, and Z, respectively, into a single file, CON_X_Y_Z. Note that the boundary points between X and Y and between Y and Z are known and saved for the file CON_X_Y_Z. Hence, we can find the corresponding boundary points in the file SIG_X_Y_Z by aligning SIG_X_Y_Z with CON_X_Y_Z in terms of a dynamic programming based algorithm. Nevertheless, the boundary points found in the file SIG_X_Y_Z are usually incorrect, and the position error between the aligned boundary and the true boundary may exceed 200ms sometimes. Therefore, manual checking is required.

III. SYNTHESIS UNIT SELECTION

In this paper, synthesis-unit selection is performed for each sentence to be synthesized as a batch, and a DP (dynamic programming) based algorithm is developed to select a globally best sequence of synthesis units for a sentence. For the DP algorithm, we defined an acoustic spectral distance function to measure the transitional smooth level of two adjacent synthesis units' formant trajectories. In addition, we derived some constrained-selection rules according to articulatory knowledge of phonemes. Some synthesis units cannot be connected because articulatory discontinuity will occur at their boundary, which may not be detected faithfully with the spectral distance function.

3.1 Spectral Distance Function

Some distance functions to measure spectral similarity between two synthesis units have been proposed in previous studies [12, 13]. Here, we modified their functions to design a distance function to meet our needs. In detail, 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

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

In this equation, $D_b(s_i, s_{i+1})$ is the first distance item for matching the boundary frames between s_i and s_{i+1} and $D_c(s_i, s_{i+1})$ is the second distance item for matching the middle frames of s_i and s_{i+1}. As to w_b and w_c, they are weighting factors that are empirically 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, respectively,

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

$$D_c = \sum_{4 \le n \le 6} d(f_n, g_n) \quad . \tag{3}$$

In Equations (2) and (3), 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 (mel frequency cepstrum coefficient) feature vectors of f_n and g_k. Note the value range of the index variable, n, in f_n is from 1 to 9. Index values from 1 to 3 mean that the frames f_n are taken from the left boundary of s_i with a frame shift of 10ms, index values from 4 to 6 mean the frames are taken from around the syllable middle-point with a frame shift of 10ms, and index values from 7 to 9 mean the frames are taken from the right boundary with a frame shift of 10ms. Similarly, the value range of n in g_n is defined in the same manner and is for the unit s_{i+1}. According to the definitions of the index values for n, it can be seen from Equation (2) that we pair f_7 with g_1, f_8 with g_2, and f_9 with g_3 instead of pairing f_7 with g_3, f_8 with g_2, and f_9 with g_1. This is because we intend to have a constant time difference between each pair of frames. Then, the variation of the distances measured from the three pairs would be smaller, and $D_b(s_i, s_{i+1})$ would more faithfully reflect the true spectral distance.

3.2 Best Sequence Selection with DP

Suppose the sentence to be synthesized is comprised of N syllables. Then, the number of possible synthesis-unit sequences that can be composed is 24^N. Therefore, we cannot search for the best sequence in an exhaustive manner but must develop a DP based algorithm to reduce the time complexity. First, define $E(i, j)$ as the minimum accumulated distance among all sequences that will stop at the j-th candidate synthesis unit of the i-th syllable but may come from a different synthesis unit for the syllables preceding the i-th syllable. Then, the recurrence relation below can be obtained according to the definition of $E(i, j)$.

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

Here, $s_{i,j}$ represents 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

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

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

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

To back track the sequence of synthesis units that accumulates to the minimum distance computed in Equation (5), we must use additional variables, *e.g.* $R(i, j)$ here, to record which value of the index k in Equation (4) will accumulate to the minimum value that is saved in $E(i, j)$. Therefore, the value of $R(i, j)$ is defined as

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

In terms of the tracking data 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 The Constrained-Selection Rules

If only spectral distances are used in the DP based best sequence searching algorithm, sometimes the synthesized speech will still be perceived to have abrupt changes (*i.e.* influent) at some syllable boundaries. This is thought to be caused by discontinuous articulator motion around syllable boundaries. Therefore, we decided to integrate articulatory knowledge of phonemes into the DP algorithm to guide the search for the best synthesis-unit sequence. In practice, we set up three constrained-selection rules to prevent two synthesis units coming from contradictory contexts from being connected into a sequence.

(A) Sentence-Start Rule

If a synthesis unit of coarticulation effect in its leading phoneme is used to synthesize the starting syllable of a sentence, the synthesized speech will be perceived as not having started from silence and as having strange articulator motion. Note that, among the four context-preceding syllables, /a, i, u, an/, the first three are single vowels and may easily be co-articulated 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 of a sentence must be extracted from a context of the form, **an_yy_zz**. Here, **zz** may be any of the 6 possible context-succeeding syllables.*

(B) Sentence-End Rule

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

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

(C) Articulator-Motion Rule

First, according to the sizes of mouth opening, vowels and consonants are divided into three classes, open-mouth phonemes, mid-mouth phonemes, and closed-mouth phonemes. A syllable is defined as starting or ending open-mouth if it starts or ends with one of the phonemes /a, o/. If a syllable starts or ends with the phoneme /i/ or a consonant, it is defined as starting or ending closed-mouth. For the other cases, a syllable is defined as starting or ending mid-mouth. In terms of the mouth-opening sizes defined above, we derived checking rules to prevent articulatory discontinuities from occurring. Suppose the two synthesis units s_{i-1} and s_i are to be connected. The rules derived are to check whether the context-succeeding syllable in the context of s_{i-1} starts open-mouth (closed-mouth) as the context-preceding syllable in the context of s_i ends closed-mouth (open-mouth). If any of the conditions checked occurs, discontinuous articulator motion is then detected. When discontinuous articulator motion is detected, the connecting of the two synthesis units, s_{i-1} and s_i, is prohibited. It is apparent that changing mouth size abruptly from open (closed) into closed (open) is a discontinuous articulator motion that is impossible to occur in real speech production.

A more practical example is shown in Fig. 3, which shows the possible connection links between the

synthesis units for the short sentence, /tai wan/. According to the sentence-start rule, the candidate synthesis units for /tai/ must be selected from the contexts that include /an/ as their context-preceding syllable. Similarly, according to the sentence-end rule, the candidate synthesis-units for /wan/ must be selected from the contexts that include /ba/ as their context-succeeding syllable. In addition, consider the synthesis unit for /wan/ that comes from the context /a_wan_ba/. Since this unit's context-preceding syllable /a/ ends open-mouth, this unit cannot be connected to a former syllable's synthesis unit that has a closed-mouth starting context-succeeding syllable such as /tai/ from /an_tai_i/ or /an_tai_ba/.

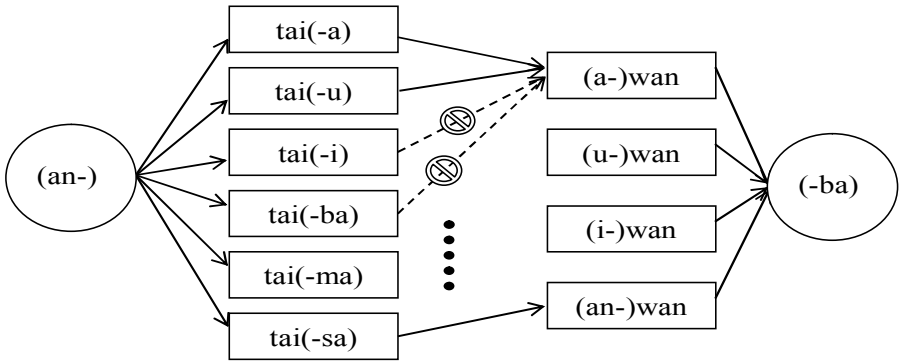

Fig. 3. Example of prohibited connections between two adjacent syllables' synthesis units.

In practical implementation, each of the constrained-selection rules given above – when detected – can be converted to a very large distance and added to the distance, $C(s_i, s_{i+1})$, which is defined in Equation (1). Then, the DP based best sequence searching algorithm discussed in Section 3.2 can still be applied. That is, the articulatory knowledge is now utilized in addition to the acoustic knowledge.

IV. SYNTHESIS EXPERIMENT AND PERCEPTUAL EVALUATION

4.1 Synthesis Experiment

To evaluate the performance of our synthesis-unit selection method, we built a Mandarin speech synthesis system, where the main processing flow is shown in Fig. 4. In the training phase, the center syllable of each of the 9,816 trisyllable contexts was extracted and fed to the HNM (harmonic-plus-noise model) [19] and to the MFCC analysis modules. Then, the obtained HNM parameters and MFCC were saved into the center syllable's corresponding parameter files. In the synthesis phase, each sentence from the input text is parsed and analyzed first. Then, the information of syllable pronunciation is fed to the DP based unit selection and to the prosody parameter generation modules. In terms of the syllable units selected and the prosody parameters generated, HNMES (harmonic-plus-noise model based and extended scheme) [20] is used to synthesize speech signals. For the generation of prosody parameters, SPC-HMM [6] is used to generate syllable pitch contours, and the other parameters are generated with rules [21].

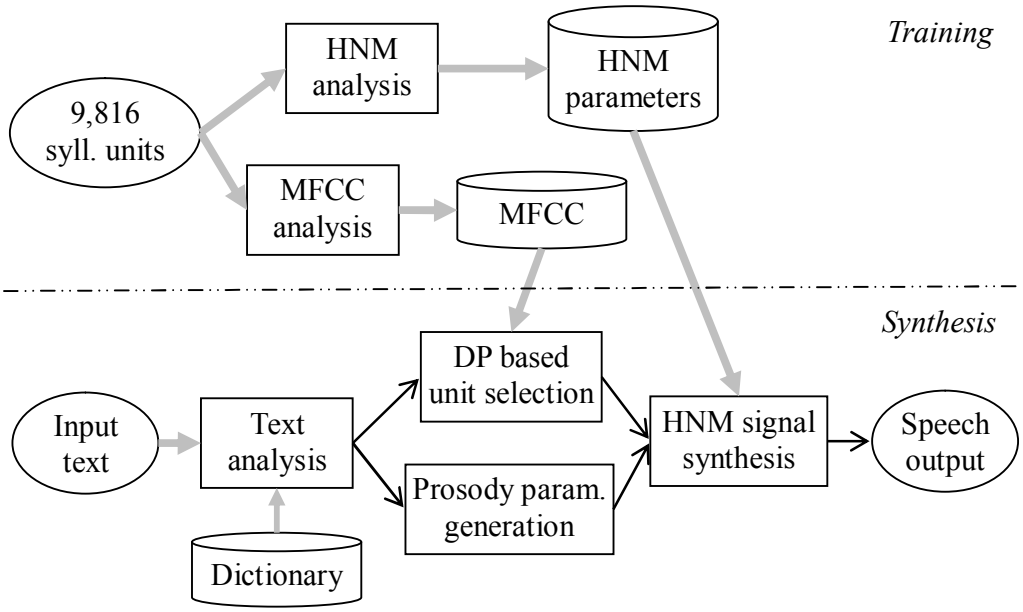

Fig. 4. Main processing flow for our Mandarin speech synthesis system.

Here, we still take the short sentence, "台灣科技," as an example input for convenience of comparison. By using the built system, the example sentence was first synthesized to obtain a speech signal. Then, the speech signal was analyzed, and its spectrogram is as shown in Fig. 5. From Figure 5, it can be seen that the F2 trajectory of /tai-2/ will go down to approach the F2 trajectory of /wan-1/, and the F2 trajectory of /ke-1/ will go up to approach the F2 trajectory of /zi-4/. These phenomena are familiarly seen in the spectrogram analyzed from a real person's uttered speech of /tai-2 wan-1 ke-1 zi-4/, as shown in Fig. 2. Therefore, the synthesis-unit selection method proposed here can indeed have the formant trajectories' transitions becoming smoother at syllable boundaries.

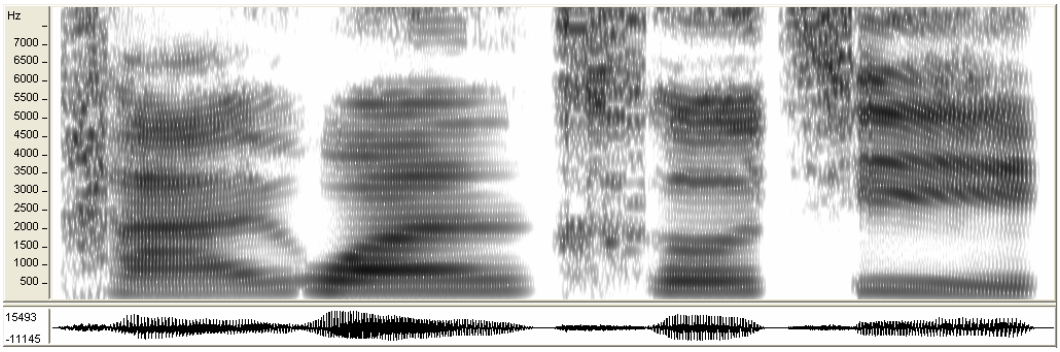

Fig. 5. Spectrogram of a speech synthesized with our unit selection method.

4.2 Perceptual Evaluation

We invited 20 persons to participate the subjective evaluation of fluency level for the synthetic speech files. Here, four speech files were synthesized from the same article of 177 Chinese characters under different synthesis conditions. These synthetic speech files are denoted as AS, AF, BS, and BF, respectively. AS and AF were synthesized under the single-unit mode (MSU). Under MSU, the synthesis unit for a syllable, yy, is always extracted from the fixed trisyllable context of the form /an_yy_ba/. In

contrast, BS and BF were synthesized under the multi-unit mode (MMU). Under MMU, the synthesis-unit selection method explained in Section 3 is used. In addition, the influence of speaking rate on fluency was also intended to be tested. Therefore, a slower speaking rate, 3.03 syllables per second on average, was adopted to synthesize AS and BS. On the other hand, a faster speaking rate, 3.85 syllables per second on average, was adopted to synthesize AF and BF. The speaking rate is controlled by adjusting a particular parameter within the prosody parameter generation module of Fig. 4. For demonstration, we have set up a web page, http://guhy.csie.ntust.edu.tw/gwtts/DynmUnit.html, from which the four synthetic speech files can be downloaded.

In the first run of listening tests, the two speech files synthesized with a slower speaking rate, AS and BS, were played in order to each of the participants. Then, each participant was requested to do a fluency comparison, *i.e.* comparing BS with AS and giving a relative fluency score. The definitions of the allowed scores had been explained to the participants beforehand. In detail, the score 3 (or -3) means that BS is much better (or worse) in fluency than AS, 2 (or -2) means that BS is better (or worse) in fluency than AS, 1 (or -1) means that BS is slightly better (or worse) in fluency than AS, and 0 means that BS and AS cannot be distinguished in fluency level. Similarly, in the second run of listening tests, the two speech files synthesized with a faster speaking rate, AF and BF, were played in order to each of the participants. Then, each participant was requested to give a relative fluency score comparing BF with AF. The definitions of the allowed scores are the same as the definitions used in the first run.

In the first run of listening tests, after collecting and analyzing the scores given by the participants, we obtained an average score of 1.175, and a standard deviation of 0.926. These are as listed in the first column of Table 1. Similarly, by collecting and analyzing the scores given by the participants in the second run of listening tests, we obtained an average score of 1.200 and a standard deviation of 1.198. These are listed in the second column of Table 1. According to the two average scores, 1.175 and 1.200, it can be seen that the fluency level will be slightly improved when the unit selection mode proposed here, MMU, is adopted instead of the reference unit selection mode, MSU. In addition, the two speaking rates, 3.03 syllables/second and 3.85 syllables/second, seem to have no influence on fluency perception since 1.175 and 1.200 are very close. Nevertheless, our opinion is that the fluency score for comparing BF with AF (both synthesized in faster speaking rate) should be significantly higher than the score for comparing BS with AS. The inconsistency between our opinion and the score, 1.200, obtained from the second run of listening tests probably is due to one factor. That is, most of the participants, 15 persons, were not familiar with the research field of speech processing, and none of the participants was familiar with speech synthesis. As a result, the scores given by the participants seem to be conservative.

Table 1. Results of subjective fluency comparisons

	BS vs. AS (slower speaking rate)	BF vs. AF (faster speaking rate)
Average score	1.175	1.200
Standard deviation	0.926	1.198

V. CONCLUSION

The discontinuities of formant trajectories at syllable boundaries are thought to be an important factor that may result in lowered acoustic fluency for synthetic Mandarin speech. Therefore, we studied and proposed a synthesis-unit selection method to have formant trajectories become smoother at syllable boundaries. This method integrates acoustic knowledge (*i.e.* the spectral distance measure) and articulatory knowledge (*i.e.* the constrained-selection rules) into the developed DP algorithm. Then, a global best sequence of synthesis units can be searched in a time-efficient way. According to the results of

the listening tests conducted, the method proposed here can indeed improve the fluency level of synthesized Mandarin speech. In detail, our unit selection method can obtain the relative fluency scores of 1.175 and 1.200 under the slower and faster speaking rate conditions, respectively, when compared with the baseline method that uses a fixed synthesis unit for each of the Mandarin syllables.

REFERENCE

[1] C. Shih and R. Sproat, "Issues in text-to-speech conversion for Mandarin," *International Journal of Computational Linguistics and Chinese Language Processing*, vol. 1, no. 1, pp. 37-86, August 1996.

[2] L. S. Lee, 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, no. 3, pp. 287-294, July 1993.

[3] S. H. Chen, S. H. Hwang, and Y. R. Wang, "An RNN-based prosodic information synthesizer for Mandarin text-to-speech", *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 3, pp. 226-239, May 1998.

[4] C. H. Wu and J. H. Chen, "Automatic generation of synthesis units and prosodic information for Chinese concatenative synthesis," *Speech Communication*, vol. 35, no. 3, pp. 219-237, Oct. 2001.

[5] G. P. Chen, G. Bailly, Q. F. Liu, and R. H. Wang, "A superposed prosodic model for Chinese text-to-speech synthesis," *International Symposium on Chinese Spoken Language Processing* (ISCSLP 2004), Hong Kong, Dec. 2004, pp. 177-180.

[6] H. Y. Gu 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, Oct. 2000, pp. 125-128.

[7] M. S. Yu, 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, Aug. 2002, pp. 21-24.

[8] C. T. Lin, 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, Feb. 2004, pp. 309-324.

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

[10] H. R. Pfitzinger, "DFW-based spectral smoothing for concatenative speech synthesis," *Int. Conf. Spoken Language Processing*, Jeju, Korea, Oct. 2004, pp. 1397-1400.

[11] F. C. Chou, "Corpus-based technologies for Chinese text-to-speech synthesis," Ph.D. Dissertation, Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, Jan. 1999.

[12] J. H. Chen, "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, June 1998.

[13] T. Toda, H. Kawai, M. Tsuzaki, and K. Shikano, "Unit selection algorithm for Japanese speech synthesis based on both phoneme unit and diphone unit," *Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 1, Orlando, USA, May 2002, pp. 465-468.

[14] M. Chu, H. Peng, H. Y. Yang, and E. Chang, "Selecting non-uniform units from a very large corpus for concatenative speech synthesizer," *Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 2, Salt Lake City, USA, May 2001, pp. 785-788.

[15] Z. H. Ling and R. H. Wang, "HMM-based hierarchical unit selection combining Kullback-Leibler divergence with likelihood criterion," *Int. Conf. Acoustics, Speech, and Signal Processing*, vol. IV, Honolulu, USA, April 2007, pp. 1245-1248.

[16] Y. Sagisaka, *et al.*, "ATR v-talk speech synthesis system," *Int. Conf. Spoken Language Processing*, Banff, Alberta, Canada, Oct. 1992, pp. 483-486.

[17] H. Y. Gu, Y. Z. Zhou, and H. L. Liau, "A system framework for integrated synthesis of Mandarin,

Min-nan, and Hakka Speech," *International Journal of Computational Linguistics and Chinese Language Processing*, vol. 12, no. 4, pp. 371-390, Dec. 2007.
[18] Z. W. Huang, "Coarticulation of two-syllable words in Mandarin speech synthesis," Master Thesis, Department of Applied Mathematics, National Chung Hsing University, Taichung, Taiwan, 1996. (in Chinese)
[19] Y. Stylianou, "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification," Ph.D. Dissertation, Ecole Nationale Supèrieure des Télécommunications, Paris, France, 1996.
[20] H. Y. Gu and Y. Z. Zhou, "An HNM based scheme for synthesizing Mandarin syllable signal," *International Journal of Computational Linguistics and Chinese Language Processing*, vol. 13, no. 3, pp. 327-342, Sept. 2008.
[21] L. S. Lee, C. Y. Tseng, and M. Ouh-Young, "The synthesis rules in a Chinese text-to-speech system," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 9, pp. 1309-1320, Sept. 1989.