Model Spectrum-progression with DTW and ANN for Speech Synthesis

Hung-Yan Gu and Chang-Yi Wu
Dept. CSIE, National Taiwan University of Science and Technology
43 Keelung Road, Section 4
Taipei, 106 Taiwan

***Abstract*-In this paper, an ANN based spectrum-progression model (SPM) is proposed. This model is intended to improve the fluency level of synthetic Mandarin speech under the situation that only a small training corpus is available. In constructing this model, first each target syllable is matched with its reference syllable by using DTW. Then, each warped path, *i.e.* spectrum-progression path, is time normalized to fixed dimensions, and used to train an ANN based SPM. After training, the SPM is used together with other modules such as text analysis, prosody parameter generation, and signal sample generation to synthesize Mandarin speech. Then, the synthetic speech is used to conduct perception tests. The test results show that the SPM proposed here can indeed improve the fluency level noticeably.**

I. INTRODUCTION

There are many languages, *e.g.* autochthons' languages and Hakka in Taiwan, that face the crisis of disappearance. It would be helpful for these languages if synthetic speech's quality can be promoted by using only a small training corpus. Note that resource (manpower and money) are usually very limited for these languages to record utterances and prepare training corpus. Therefore, we keep in mind that speech synthesis techniques developed should be economically transferrable to another language. Even though the language studied here is Mandarin and is of large population, we still study to improve its synthetic speech's quality by using only a small training corpus.

Speech-Fluency Factors: FT Discontinuity vs. Spectrum Progression

It is well known that prosody models play important roles in synthesizing natural Mandarin speech [1, 2, 3]. The prosodic parameters that are modeled include pitch contour, syllable duration, amplitude, and syllable-front pause. In the past, we had some research experiences in model-based speech synthesis, *i.e.* prosodic parameters and signal waveforms are separately modeled and generated. However, the synthetic speech is not perceived as fluent as uttered by a real person even thought the generated prosodic parameters are natural enough. We once attributed this phenomenon to the discontinuities of formant traces (FT) between two adjacent syllables. Therefore, we had spent some efforts to solve the problem of formant trace discontinuities by means of unit selection [4]. Some improvement in fluency is observed when adopting this method. Nevertheless, a gap in fluency is still perceived between synthetic and real speeches.

Recently, we found that the problem, lack of fluency, is already noted by several researchers [5, 6, 7]. The solution method they proposed is to slice a syllable into several segments in terms of an optimal state sequence of an HMM (hidden Markov model). Each segment is characterized by a specific spectral envelope that is modeled with a Gaussian mixture on a state. In addition, the duration of a segment is modeled with a state-staying PDF (probability density function). In our viewpoint, the HMM based method is a more detailed planning method to subdivide a syllable's duration into several comprising segments' durations in a non-uniform manner. That is, different segments of different spectral envelope are assigned unequal durations. The purpose of such unequal assignments of segment durations we think is to more delicately mimic the progression of spectrum with time in which a real person articulates.

Spectrum-Progression Path and Modeling

The dynamic changing of spectrum (or spectral envelope) with time is an important concept, and is termed "spectrum progression" here. Also, the term, spectrum progression path (SPP), is frequently used. SPP means a spectrum mapping curve between a syllable uttered within a sentence and a syllable uttered in isolation that has same pronunciation. An example of SPP is shown in Fig. 1, in which the waveform on

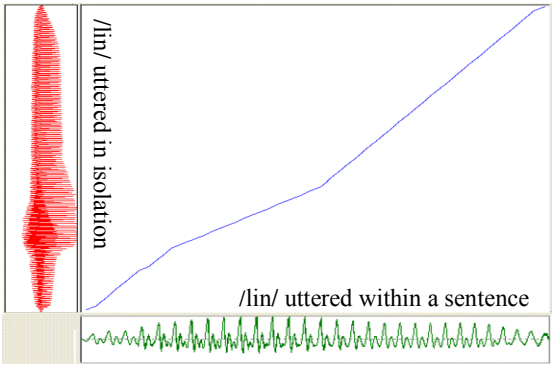

Fig. 1. An example of spectrum progression path.

the horizontal axis is the syllable /lin/ uttered within a sentence and the waveform on the vertical axis is the same syllable but uttered in isolation. In the past, synthetic speeches from many Mandarin speech synthesis systems are felt as lacking of fluency. We think the main cause is that they don't adopt a SPP model and set the SPP directly to a straight line.

Therefore, we began to study the problem of SPP modeling and generation. Here, we don't follow previous researchers to adopt HMM. One major reason is that our training corpus is small and is not affordable for training HMM. Another minor reason is that the assumption of HMM, a state's state-staying

duration PDF is independent of its adjacent states, seems unreasonable. Also, another minor reason is that we wish the acoustic parameters used in generating speech waveform need not be restricted to the same parameters for representing spectral envelope. Actually, in this paper, we use MFCC (Mel frequency cepstrum coefficient) [8, 9] for SPP searching but use HNM (harmonic plus noise model) parameters for generating speech waveform. Because of these reasons, we decide to adopt ANN (artificial neural network) to model spectrum progression, and such ANN is thus called ANN spectrum progression model (SPM).

II. SPECTRUM PROGRESSION MODEL

Before SPM can be used to generate SPP, it must be trained first. The training procedure consists of the processing steps drawn in Fig. 2. In the first block, 375 training sentences and

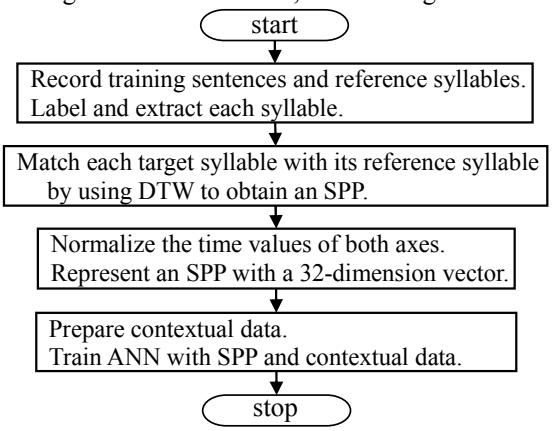

Fig. 2. SPM training procedure.

408 reference syllables are uttered by a female speaker and recorded into a computer. The number of syllables in the training sentences is totally 2,926. Then, each of the syllables comprising a training sentence is labeled and extracted. In the second block, each target syllable, *i.e.* a syllable extracted from a training sentence, is placed on the horizontal axis of Fig. 1 and matched with its reference syllable that has same pronunciation to obtain a DTW (dynamic time warping) based SPP. Here, a reference syllable is a syllable uttered in isolation, and is placed on the vertical axis of Fig. 1. Next, in the third block, the time lengths of target and reference syllables are separately normalized to one unit. Then, 32 time points are uniformly placed on the time axis of the target syllable. For each time point, a corresponding point at the time axis of the reference syllable is mapped with the SPP just obtained in the last block. Hence, a SPP can be represented with a 32-dimension vector of mapped time values. Afterward, in the bottom block of Fig. 2, contextual data are prepared for each target syllable, and used together with that target syllable's SPP data to train the ANN SPM.

A. Dynamic Time Warping

Dynamic time warping is a traditional method for speech recognition [8]. Here, we use DTW to find a best SPP between a target syllable and a reference syllable. Before applying DTW, feature vectors must be extracted first. Let $X = X_1, X_2,\ldots,X_n$, be the sequence of feature vectors extracted from a target syllable, and $Y = Y_1,Y_2,\ldots,Y_m$, be the sequence of feature vectors extracted from a reference syllable. Here, a feature vector is extracted from a Hamming windowed signal frame of length 20ms and shifted every 5ms. The elements of a feature vector are 13 MFCC coefficients and 13 delta values of MFCC [8, 9].

Because the two sequences, X and Y, are of different length in general, *i.e.* $n \neq m$, the matched path, of minimum distance, between X and Y must be computed with a dynamic programming algorithm to reduce the amount of computations required. The local continuity constraint adopted here is shown in Fig.3. Such a local constraint is adopted because it

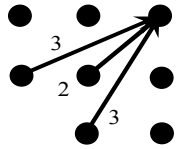

Fig. 3. Local constraint for DTW.

provides uniqueness of mapping from horizontal axis to vertical axis. According to such local constraint, the accumulated distance, $Da(X_i, X_j)$, from the origin to the grid point, (X_i, X_j), is computed recursively as

$$D_a\left(X_i,Y_j\right)=\min\begin{Bmatrix} D_a\left(X_{i-1},Y_{j-2}\right)+3\cdot D\left(X_i,Y_j\right) \\ D_a\left(X_{i-1},Y_{j-1}\right)+2\cdot D\left(X_i,Y_j\right) \\ D_a\left(X_{i-2},Y_{j-1}\right)+3\cdot D\left(X_i,Y_j\right) \end{Bmatrix} \quad (1)$$

where the distance function, $D(X_i, X_j)$, compute the geometric distance between X_i and X_j, and the local path weights, 3 and 2, are used to avoid path biasing.

When executing DTW, we find that the sequence length, n, of a target syllable is often much smaller than the sequence length, m, of its reference syllable. This would cause the endpoint, (n, m), cannot be reached when applying Equation (1). To solve this problem, the condition, $m/n < 1.5$, is checked before executing DTW. If the condition is not satisfied, the frame shift of the reference syllable is automatically adjusted to reduce the number of frames. The adjusting formula used here is

$$L_b = L_a \times \frac{2\cdot m}{3\cdot n} \quad (2)$$

where *La* denotes the length of original frame shift and *Lb* denotes the length of the adjusted frame shift.

B. ANN Structure

The structure of the ANN adopted here is shown in Fig. 4. The input layer has 28 nodes to input 8 types of contextual data. The output layer has 32 nodes to output 32 ordered spectrum progression parameters. In addition, there are one hidden layer and one recurrent hidden layer. The quantities of nodes in the two hidden layers are equal. However, the actual number of nodes to use will be determined according to the error value returned during training the ANN link-weight parameters. The training algorithm adopted here is a steepest decent method [10]. After training the ANN with different number of nodes, we find that placing 16 nodes to the hidden layers is the best according to the error values returned.

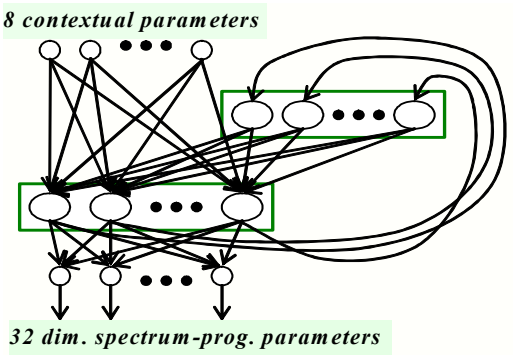

Fig. 4. ANN structure for spectrum progression modeling

As to the contextual data, consider that the spectrum progression of the t-th syllable, S_t, of a sentence is chiefly influenced by the tone type, syllable initial and final types of the syllable, S_t. In addition, the tone type and final type of the last syllable, S_{t-1}, and the tone type and initial type of the next syllable, S_{t+1}, may still have some effects. Therefore, the contextual data prepared for each target syllable are as the 8 types listed in Table 1. In Table 1, the last item, time index, is

Table 1 Contextual data for a target syllable

Items	tone of S_{t-1}	class of S_{t-1} final	tone of S_t	init. of S_t	final of S_t	tone of S_{t+1}	class of S_{t+1} init.	time Index
Bits	3	4	3	5	6	3	3	Void

intended to carry timing information. If the current syllable is the t-th syllable of a sentence of T syllables, then the value of time index is set to the floating-point number t/T. For the data items of tone, 3 bits are allocated to represent each of the tone items because Mandarin has 5 different tones. For the initial of St, 5 bits are allocated because Mandarin has 21 different syllable initials. Similarly, 6 bits are allocated to represent the final of St because Mandarin has 39 different syllable finals. Here, because the number of recorded target syllables is not large enough, we decide to group the final types of St-1 into 9 classes and group the initial types of St+1 into 6 classes in order to decrease the number of possible combinations for the ANN's input data. Hence, just 4 bits and 3 bits are allocated to represent these classes, respectively.

III. Integration of Our System

After the ANN SPM is trained, we integrate it with other modules to build a Mandarin speech synthesis system. With this system, wave files for perception tests are then synthesized. The other modules involved are as those blocks in Fig. 5, including text analysis, prosodic parameter generation, and signal waveform synthesis. When the text of a Chinese sentence is inputted, it is parsed, in the "text analysis" block, into a sequence of words by looking up word dictionaries. In the same time when dictionaries are looked up, a Chinese character's pronunciation syllable and tone are determined. Next, tone sandhi rules are applied. Then, the sequence of syllables and tones are fed to both the left and right blocks of Fig. 5 to generate spectrum progression parameters and prosodic parameters for each syllable.

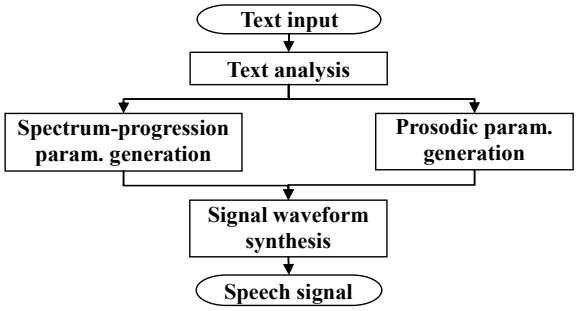

Fig. 5. System structure of our Mandarin synthesis system

A. Generation of Prosodic Parameters

The prosodic parameters for each syllable include pitch contour, duration, and intensity. These parameter values are determined in the right block of Fig. 5. Here, we have constructed a separate ANN for each of the three prosodic parameters, and each ANN is of the structure as shown in Fig. 4. In fact, the ANN for pitch contour generation was trained in our previous work [11] by using 375 training sentences uttered by a male speaker. The reason for not building a combined ANN for the three parameters is that the quantity, 375, of training sentences is not large enough. However, by appropriate grouping the values of some contextual data items as seen in Table 1, the generated prosodic parameters can still show acceptable performance in naturalness level. This can be checked by browsing our web page, http://guhy.csie.ntust.edu.tw/spmdtw/, to listen to the example wave files of synthetic speech.

B. Synthesis of Signal Waveform

In the bottom block of Fig. 5, signal samples for each syllable, St, are synthesized in terms of its reference syllable. Because each reference syllable is recorded and saved only once, hence there is no chance to do unit selection. In addition, consider that syllable signals of diverse prosodic characteristics need to be synthesized, and the signal waveforms synthesized by the traditional method, PSOLA [12], are not stable in quality when pitch contour or duration is considerably changed. Therefore, we decided to develop an HNM (harmonic plus noise model) [13] based method for synthesizing the signal samples of a Mandarin syllable. By the efforts spent in the previous study [14], we had developed an HNM based and improved method that satisfies our needs.

Before the synthesis procedure can start, each reference syllable must be analyzed beforehand to obtain its HNM parameters, *i.e.* amplitude, frequency, and phase parameters for each harmonic partial and 20 cepstrum coefficients for representing the envelope of noise spectrum. In our synthesis procedure, the first step is to place control points uniformly on the time axis of a synthetic syllable. The spacing between two adjacent control points is fixed to 100 sample points. As the second step, the HNM parameters for each control point are derived. By using the spectrum progression parameters generated in the left block of Fig. 5, the time position of a control point can be mapped to a time point on the reference syllable's time axis. Accordingly, two adjacent frames surrounding this time point are located, and the HNM

parameters of the two frames are used to interpolate out the HNM parameters for the control point [14]. As the third step, the frequencies of harmonic partials on a control point are adjusted to have a pitch height that meets the pitch contour generated by the right block of Fig. 5. When the frequency of a harmonic partial is changed, its amplitude must also be adjusted in order to keep timbre consistent [14]. Then, as the final step, signal samples' values are computed in terms of the adjusted HNM parameters on two adjacent control points.

IV. Synthetic SPP and Perception Test

A. Synthetic Spectrum Progression Path

To have an impression of synthetic SPP, here we take the training sentence, "cing3 ba3 zhe4 lan2" (Please handle the basket), as an example, and let the SPM ANN (as shown in Fig. 4) generate an SPP for each syllable of the sentence. As a result, the SPP for the four syllables are as those drawn in Fig. 6. The numbers on the horizontal axis index the dimensions of a synthetic SPP represented as a 32-dimension vector. The synthetic SPP for the two syllables, /ba3/ and /lan2/, would go above the middle line (form bottom left to top right) while the SPP for the other two syllables would go below the middle line. These trends of SPP movements are consistent with the SPP obtained from matching the corresponding recorded syllables with their reference syllables.

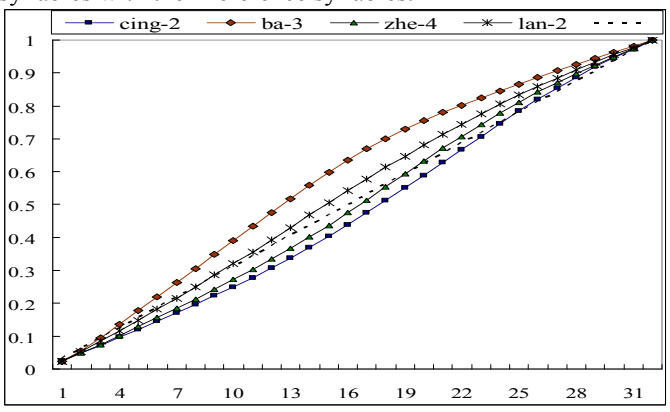

Fig. 6. Synthetic SPP for four syllables

B. Perception Test

A short article of about 130 syllables is fed to our Mandarin synthesis system. At the first time, the left block of Fig. 5 is disabled and spectrum progression parameters that represent linear time mapping are generated. Under such condition, the synthetic speech file is denoted as *VA*. In contrast, the left block of Fig. 5 is enabled at the second time, and the synthetic speech file obtained is denoted as *VB*. Then, we play *VA* and *VB* in order to each of 15 participants, respectively. Each participant is requested to give a score about the fluency difference of *VB* versus *VA*. A score of 2 (-2) means *VB* (*VA*) is apparently more fluent than *VA* (*VB*). A score of 1 (-1) means *VB* (*VA*) is slightly more fluent than *VA* (*VB*). A score of 0 means the fluency level cannot be distinguished. After perception tests, the given scores are analyzed. The average score obtained is 1.33 while the standard deviation is 0.471. The average, 1.33, indicates that the use of SPM built here can indeed promote the fluency level. The interested reader can browse the web page, http://guhy.csie.ntust.edu.tw/spmdtw/, to download the synthetic voice files.

V. Conclusion

In this paper, an SPM that makes use of DTW and ANN is studied. A target syllable is matched with its reference syllable by using DTW to obtain an SPP. Then, the SPP of all target syllables are used to train an ANN SPM. By integrating the SPM and other modules, we built a Mandarin speech synthesis system. The files of synthetic speech are then used to conduct perception tests. According to the results of the tests, the SPM proposed here can indeed promote the fluency level of the synthetic Mandarin speech significantly. Therefore, the SPM proposed here can be viewed as an alternative to HMM for modeling the progression of spectrum with a small training corpus. In addition, according to perception testing, modeling the progression of spectrum within a syllable can obtain much more fluency-level improvement than solving the discontinuous movements of formant traces between two adjacent syllables.

References

[1] 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, 1998.

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

[3] 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*, Taipei, Taiwan, pp. 21-24, 2002.

[4] H. Y. Gu and K. H. Wang, "An Acoustic and Articulatory Knowledge Integrated Method for Improving Synthetic Mandarin Speech's Fluency", *International Symposium on Chinese Spoken Language Processing*, Hong Kong, pp. 205-208, 2004.

[5] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, "Duration Modeling in HMM-based Speech Synthesis System", *International Conference on Spoken Language Processing*, Vol. 2, pp. 29–32, 1998.

[6] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based Speech Synthesis System Applied to English", *IEEE Workshop on Speech Synthesis*, Santa Monica, CA, pp. 227-230, 2002.

[7] Y. Qian, F. Soong, Y. Chen, and M. Chu, "An HMM-Based Mandarin Chinese Text-to-Speech System", *International Symposium on Chinese Spoken Language Processing*, Singapore, Vol. 1, pp. 223-232, 2006.

[8] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, 1993.

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

[10] K. Gurney, *An Introduction to Neural Networks*, UCL Press, 1997.

[11] 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, 2007.

[12] E. Moulines and E Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", *Speech Communication*, Vol. 9, pp. 453-467, Dec. 1990.

[13] Yannis Stylianou, *Harmonic plus Noise Models for Speech, Combined with Statistical Methods, for Speech and Speaker Modification*, Ph.D. thesis, Ecole Nationale Supèrieure des Télécommunications, Paris, France, 1996.

[14] H. Y. Gu and Y. Z. Zhou, "Mandarin Syllable Signal Synthesis Using an HNM Based Scheme", *International Conference on Audio, Language, and Image Processing*, Shanghai, China, pp. 1635-1639, 2008.