Combining HMM Spectrum Models and ANN Prosody Models for Speech Synthesis of Syllable Prominent Languages

Hung-Yan GU, Ming-Yen LAI and Sung-Feng TSAI

Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology, Taipei
e-mail: {guhy, M9615074, M9615069}@mail.ntust.edu.tw

Abstract—**In this paper, an approach that combines HMM spectrum models and ANN prosody models is proposed to construct a speech synthesis system. Currently, a Mandarin corpus is used to show the feasibility of this approach. We hope that this approach can be used in other syllable prominent languages like Min-Nan and Hakka. In the training phase, DCC (discrete cepstrum coefficients) are computed for each frame of the training corpus and used as spectral parameters. Multiple utterances of a syllable are first grouped into a few clusters according to their DTW paths. Then, each cluster's syllable utterances are used to train an HMM. In the synthesis phase, for each syllable of a sentence, an HMM of the syllable is selected first according to this syllable's contextual data. Then, a duration ANN and duration means of the HMM states are used to determine how many frames an HMM state should be assigned. To achieve the goal of real-time synthesis, we propose an interpolation method to generate DCC coefficients for each frame. Next, speech signal is synthesized by using the DCC coefficients and the pitch contour generated by another ANN to control an HNM (harmonic plus noised model) based signal synthesizer. The results of perception tests show that our interpolation method obtains slightly more natural synthetic speech than the MLE method. Also, the duration ANN can have more natural synthetic speech than the duration means of HMM states.**

Keywords-speech synthesis; spectrum model; prosody model; discrete cepstrum; HMM; ANN; HNM

I. INTRODUCTION

Mandarin, Min-Nan and Hakka are all syllable prominent languages, and have their own populations in Taiwan. We hope that a speech synthesis technique developed for a language can be economically transferred to other languages. In this paper, we use a Mandarin corpus to develop a speech synthesis technique, i.e. combining HMM spectrum models and ANN prosody models. Nevertheless, we think the developed technique can be economically transferred to Min-Nan or Hakka since economical transferability is always kept in mind.

Recently, unit selection based Mandarin speech synthesis methods were studied by several researchers to obtain a certain improvement of naturalness level [1-3]. Nevertheless, we did not adopt this approach (unit selection) in our previous studies [4-7]. One of our reasons is that a large amount of labor and time is still required to record and prepare a speech corpus for another language (e.g. Min-Nan and Hakka) that the synthesis technique is to be applied to. Also, another reason is that we hope the speech synthesis system constructed is multifunctional. For example, it should support the function of speaking-rate adjusting and the function of timbre transformation (e.g. transforming a female's timbre into a male's timbre).

To implement the functions mentioned, we therefore adopted the approach of parametric synthesis till now. Under the premise of parametric synthesis, the two major kinds of factors that influence a synthetic speech's naturalness we think are acoustic fluencies and prosodic fluencies. To improve the acoustic fluency of spectrum progression, we recently studied and proposed a method to model the spectrum progression of a syllable [7]. Such a method can indeed significantly improve the acoustic fluency of spectrum progression within a syllable. Nevertheless, the acoustic fluency of inter-syllable connection (i.e. continuities of formant trajectories) is still not satisfactory. In contrast, the approach of modeling spectrum progression with HMM was studied recently by many researchers [8-10]. Therefore, we thought HMM may be a good choice, and began to model the spectrum progression of a syllable with HMM.

On the other hand, prosodic fluencies are also very influential to a synthetic speech's naturalness. For example, the pitch contours of a sequence of syllables would be perceived as fluent if the pitch contours of the syllables are of consistent pitch heights and the pitch contours of any two adjacent syllables are kept continuously varied (i.e. no pitch discontinuity in syllable boundary). Similarly, the speaking rate of a sequence of syllables would be perceived as fluent if the durations of the syllables are of consistent lengths. In an MSD-HMM (multi space probability distribution HMM) [8, 9], a speech unit's pitch-contour is modeled with its spectrum progression simultaneously. However, we think that the modeling of pitch-contour (or syllable duration) is better considered separately. Our viewpoints include that many superior prosody-parameter generation models were already developed by previous researchers, and that MSD-HMMs are just context-dependent HMMs without explicitly utilizing the concept of prosodic-states [5] and are hence suspected to model sentence intonations and declining phenomena well.

Therefore, we decide to combine syllable spectrum-progression HMMs, a syllable pitch-contour generation ANN, a syllable duration generation ANN, and an HNM (harmonic-plus-noise model) [6, 11] based signal synthesis module to construct a speech synthesis system for Mandarin. It is hoped that such a combination will obtain not only the acoustic fluency in spectrum progression but also the prosodic fluencies in pitch contours and syllable durations. Furthermore, we hope that the system constructed can be run in real-time in a personal computer. Therefore, the bottleneck in generating signal frames' spectral parameters is also considered and eliminated.

In the following, the training of the HMM spectrum models and the ANN prosody models will be described in Section 2. In Section 3, the main processing flow for synthesizing Mandarin speech is given first. Then, the function of each block is explained. Next, perception tests and results are described in Section 4. Finally, concluding remarks are given in Section 5.

II. MODEL TRAINING

To train the models used in our system, the work flow as depicted in Fig. 1 is followed.

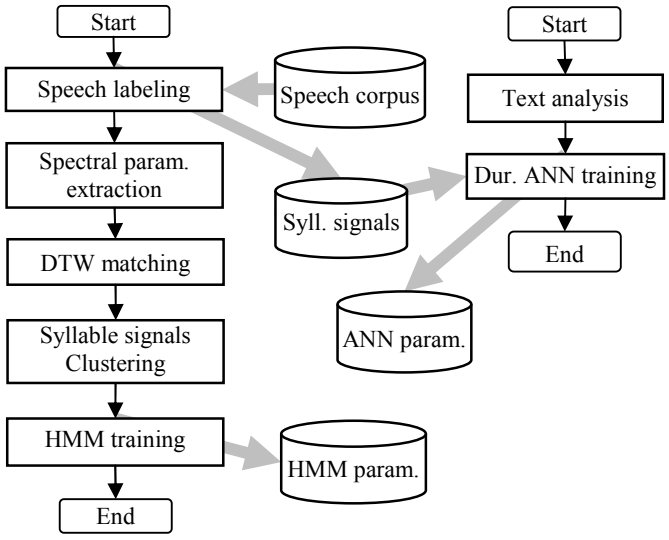

Figure 1. The work flow for training the HMM and ANN models

A. Recording and Labeling

In a sound-proof room, a male adult was invited to pronounce 1,208 Mandarin sentences. The text script consists of 1,208 independent sentences, i.e. no relation exists in adjacent sentences, and it has totally 10,173 Chinese characters. In addition, he was also requested to utter the 407 Mandarin syllables in isolation in order to use these syllable signals for DTW matching. Here, the sampling rate is 22,050Hz. For labeling syllable signals, the package, HTK, was used first to perform forced alignment. Then, the software, WaveSurfer, is used to adjust syllable boundaries.

B. Spectral Parameter Extraction

A recorded speech file is sliced into a sequence of frames. The frame width is set to 512 sample points and the frame shift is 256 points. For each frame, a vector of 39 spectral parameters, i.e. DCC coefficients [12], c_0, c_1, …, c_{38}, was extracted. The processing flow for extracting DCC is shown in Fig. 2. The details for the processing blocks are referred to a previous work [13].

C. DTW matching and Syllable Signal Clustering

The spectrum progression within a syllable is affected by the preceding and succeeding phonemes, i.e. contextual dependency. Therefore, the method of deriving a spectrum progression path (SPP) with DTW [7] is adopted here. Each syllable uttered within a sentence is placed at the X-axis while its corresponding syllable uttered in isolation is placed at the

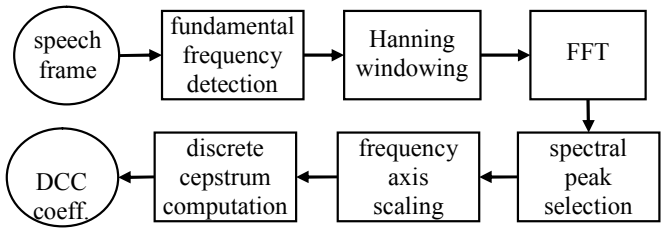

Figure 2. The processing flow for extracting DCC coefficients.

Y-axis. An SPP is represented here as a vector of 32 warped and normalized time values.

Note that a Mandarin syllable may be uttered several times within some sentences of the corpus. Hence, the SPP of the syllable signals uttered for a same Mandarin syllable (tones are not distinguished) are collected to perform K-means clustering. Then, the syllable signals of similar SPP will be clustered into a class. Here, the number of classes, *NP*, for a Mandarin syllable is determined empirically. In detail, *NP* is set by dividing the number of syllable signals with 10. After K-means clustering, each syllable signal's information is saved, e.g. the class it is clustered to, the adjacent phonemes it is surrounded.

D. HMM Spectrum Model Training

The syllable signals clustered into a class were used to train an HMM to model their spectrum progression. Here, each HMM is structured to have 8 states and transited in a left-to-right manner without state skipping. In each state, only one Gaussian mixture is adopted. For training an HMM, the algorithm of segmental K-means [14] is used here.

As mentioned in Section 2.B, 39 DCC coefficients are extracted from a speech frame. Nevertheless, the states of an HMM must be distinguished whether they are voiced or unvoiced when this HMM is used in speech synthesis processing. Therefore, the periodicity of a frame must also be saved. Here, an extra dimension, i.e. c_{39}, is used to indicate a frame's periodicity, i.e. setting $c_{39} = 1$ if voiced and setting $c_{39} = 0$ if unvoiced. In addition, since differential spectral parameters are useful, 40 more dimensions are added to represent delta DCC. After training an HMM, the average number of frames staying at a sate and its variance are also saved besides the HMM parameters.

E. ANN Prosody Model Training

To generate the prosodic parameters, syllable durations and pitch-contours, separate ANNs are used in our system. Here, the speech corpus of 10,173 syllables is used to train the duration ANN. Nevertheless, the pitch-contour ANN is an old one, i.e. it is directly taken from our previous study [7].

The duration ANN has 28 nodes in the input layer for inputting 8 contextual parameters, and one node in the output layer for outputting a syllable's duration. More details about the contextual parameters are referred to our previous work [7]. In addition, the ANN has one hidden layer and one recurrent hidden layer. The number of nodes to be placed in the hidden layers was tested from 15 to 20. The best choice was found to be 17. On the other hand, the pitch-contour ANN has also 28 nodes in the input layer to input same contextual parameters. Nevertheless, it has 16 nodes in the output layer to output 16

time-normalized pitch-frequency values to represent a syllable's pitch-contour.

III. SPEECH SYNTHESIS PROCESSING

The main processing flow of our system for synthesizing Mandarin speech signals is shown in Fig. 3. In the block of text analysis, a sentence is parsed from the text input each time. Then, the word dictionary is looked up to segment the sentence into a sequence of words and get a pronunciation syllable for each character. Next, in the block of "generate pitch-contours and durations", 8 contextual data items are prepared for each syllable of the sentence first. Then, the contextual data are fed one after another to the two ANNs to generate a pitch-contour and a duration value for each syllable. As to the other blocks, their operations will be described in the following subsections.

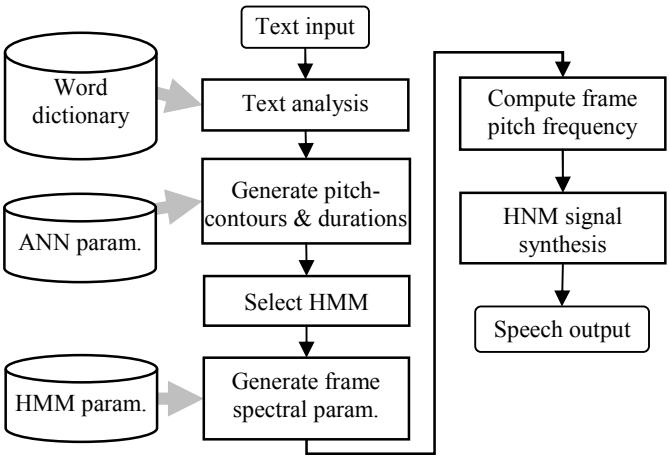

Figure 3. The main flow for synthesizing Mandarin speech signals.

A. Syllable HMM Selection

The syllable signals, uttered within different contexts for a same base syllable (with tone disregarded), were already clustered into several classes according to their SPP, and the syllable signals of each class were used to train a corresponding HMM. Then, when a syllable's signal is to be synthesized, which HMM of its corresponding base syllable should we select?

As described in Subsection 2.C, the contextual data collected from each syllable signal were sorted and saved into a search table. About the format of the search table, an example is given in Fig. 4. In Fig. 4(a), the data items listed in each row are, respectively, current syllable, final of previous syllable, initial of next syllable, initial of previous syllable, final of next syllable, and index to the HMM that the current syllable's signal was used to train. If an asterisk is used instead, it means the item is a null initial or final. When the signal of a syllable is to be synthesized, it and its contextual data are used to search the table as shown in Fig. 4(a). If an exact match is found, the last data item of the matched row, i.e. the HMM index, is used to get the right HMM. Otherwise, the table is simplified first by eliminating the minor data items, i.e. the fourth and fifth items of a row. An example simplified table is as that shown in Fig. 4(b). Then, the simplified table is searched. If still no match is found, the table is simplified furthermore and searched again.

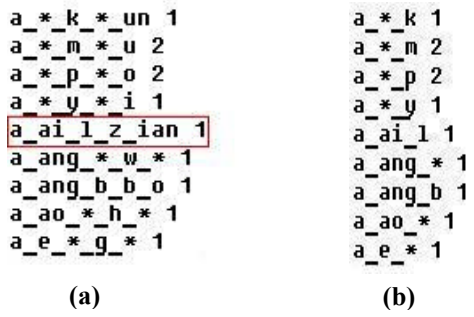

Figure 4. An example of search tables for HMM selection

B. Spectral Parameter Generation

After a syllable's right HMM is selected, the states of the HMM will be assigned some numbers of frames according to the syllable-duration value generated by the ANN. The formula used here for assigning the numbers of frames is referred to a typical HMM based speech synthesis work [9]. Next, consider how to generate each frame's spectral parameters, i.e. DCC coefficients. One commonly used method is based on MLE (Maximum likelihood Estimate) [9]. Here, in order to achieve real-time speech synthesis, we studied and proposed a faster generation method, named WLI (weighted-linear interpolation). Although this method just generates approximated spectral parameters, it can be run in 30 more times faster than the MLE method, and the perceived quality of the synthetic speech is still acceptable and comparable with that synthesized by using the MLE method.

The details of the WLI method are as the following. Suppose the DCC for the frames between states S_i and S_{i+1} are to be generated. One example situation is illustrated in Fig. 5. First, compute the difference between the two mean DCC vectors on states S_i and S_{i+1} respectively, i.e.

$$DF_j^i = cm_j^{i+1} - cm_j^i , \quad j=0, 1, 2, ..., 38 , \tag{1}$$

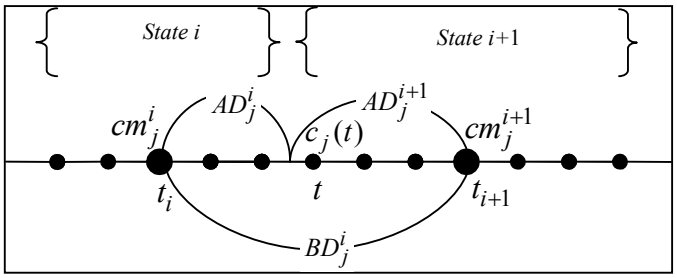

Figure 5. An example situaion used for explaining the WLI method

where cm_j^i denotes the j-th dimension of the mean DCC vector on state i. Also, compute half the difference, AD_j^i, that will be introduced by state S_i in ordinary speaking rate, i.e.

$$AD_j^i = \Delta cm_j^i \times \frac{L_i}{2} , \quad j=0, 1, 2, ..., 38 , \tag{2}$$

where Δcm_j^i is the mean delta for cm_j^i on state S_i and L_i is the number of frames assigned to S_i. Then, the distance introduced in ordinary speaking rate, BD_j^i, is computed as

$$BD_j^i = \begin{cases} AD_j^i + AD_j^{i+1} , & if \ \Delta cm_j^i \times \Delta cm_j^{i+1} \geq 0 \\ (L_i + L_{i+1}) / 2 , & if \ \Delta cm_j^i \times \Delta cm_j^{i+1} < 0 \end{cases} . \tag{3}$$

Next, the DCC for the *t*-th frame is generated as

$$c_j(t)=\begin{cases} cm_j^i+(t-t_i)\cdot\Delta cm_j^i\cdot R_j^{i\text{-}1}, & if\ \ t<t_i\ \ and\ \ \Delta cm_j^{i-1}\cdot\Delta cm_j^i\geq 0 \\ cm_j^i+(t-t_i)\cdot R_j^{i\text{-}1}, & if\ \ t<t_i\ \ and\ \ \Delta cm_j^{i-1}\cdot\Delta cm_j^i<0 \\ cm_j^i+(t-t_i)\cdot\Delta cm_j^i\cdot R_j^i, & if\ \ t\geq t_i\ \ and\ \ \Delta cm_j^i\cdot\Delta cm_j^{i+1}\geq 0 \\ cm_j^i+(t-t_i)\cdot R_j^i, & if\ \ t\geq t_i\ \ and\ \ \Delta cm_j^i\cdot\Delta cm_j^{i+1}<0 \end{cases} \quad (4)$$

where $R_j^i = DF_j^i \,/\, BD_j^i$.

C. Pitch f0 Computation and Signal Synthesis

When a syllable's signal is to be synthesized, the unvoiced boundary state of its selected HMM must be decided first if it has an unvoiced initial phoneme. This can be done by checking the parameter, cm_{39}^i, for i = 0, 1, …, 7. If cm_{39}^i is less than 0.5, then the i-th state is decided to be unvoiced. For those states decided to be voiced, their assigned frame numbers are accumulated to obtain a total number of voiced frames. Then, the 16 pitch-contour parameters generated by the ANN are Lagrange interpolated to compute an f0 value for each voiced frame.

When the DCC and f0 for each frame of a syllable are generated, they are then fed in frame order to the block "HNM signal synthesis" in Fig. 3 to synthesize speech signals. The details for signal synthesis are referred to a previous work [13].

IV. Synthesis Experiments and Perception Tests

After the system was constructed, its synthesis speed was tested on a notebook computer with an Intel T5600 1.83GHz CPU. As a result, it spends 29.1 sec. to synthesize a speech file of 49.2 sec. in length. Therefore, our system can indeed be run in real-time.

For conducting perception tests, we used the system to synthesize three speech files, SA, SB, and SC under different settings. SA is synthesized by using the WLI method to generate DCC while SB is synthesized by using the MLE method. As to SC, it is also synthesized by using the WLI method but syllable durations are just according to the duration means of HMM states. Here, 15 persons were invited to participate two runs of perception tests. In the first run, SA and SB are played to each participant, and he (or she) is requested to give a score about comparing SB with SA. The defined value range for a score is from -2 to 2. A positive score means SB is better than SA in naturalness. As a result, the average score obtained is -0.528, which indicates SA is slightly more natural than SB and the WLI method is acceptable in its synthetic-speech quality. In the second run, SA and SC are played to each participant to compare their naturalness, and he (or she) is requested to give a score. The average score obtained for this run is -1.03, which indicates SA is better than SC in naturalness and the syllable durations generated by the ANN are better than those directly taken from HMM state duration means. For demonstration, we have prepared a web page to provide example synthetic speech files for listening, i.e. http://guhy.csie.ntust.edu.tw/hmmsyn/ .

V. Concluding Remarks

In this paper, a speech synthesis approach that combines HMM spectrum models and ANN prosody models is proposed. The meaning of such combination is that previously developed superior prosody models can still be used to help HMM based spectrum progression models to simultaneously promote both kinds of fluencies, i.e. prosodic and acoustic fluencies. To show the feasibility of this approach, we used a Mandarin corpus to build a real-time speech synthesis system for Mandarin. We think the technique developed can be economically transferred to other syllable prominent languages, e.g. Min-Nan and Hakka. In addition, although the WLI method for generating spectral parameters is proposed to achieve real-time speech synthesis, its synthetic-speech quality is however comparable with that synthesized by using the MLE method.

Acknowledgment

This study is supported by National Science Council of Taiwan under the contract number, NSC 98-2221-E-011-116.

References

[1] F. C. Chou, *Corpus-based Technologies for Chinese Text-to-speech Synthesis*, Ph.D. Dissertation, Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, 1999.

[2] M. Chu, H. Peng, H. Y. Yang, and E. Chang, "Selecting Non-uniform Units from a Very Large Corpus for Concatenative Speech Synthesizer", *ICASSP*, Salt Lake City, USA, pp. 785-788, 2001.

[3] Z. H. Ling and R. H. Wang, "HMM-based Hierarchical Unit Selection Combining Kullback-Leibler Divergence with Likelihood Criterion", *ICASSP*, vol. IV, Honolulu, USA, pp. 1245-1248, 2007.

[4] H. Y. Gu and W. L. Shiu, "A Mandarin-syllable Signal Synthesis Method with Increased Flexibility in Duration, Tone and Timbre Control", *Proc. Natl. Sci. Counc. ROC(A)*, Vol. 22, pp. 385-395, 1998.

[5] H. Y. Gu and C. C. Yang, "A Sentence-pitch-contour Generation Method Using VQ/HMM for Mandarin Text-to-speech", *ISCSLP*, Beijing, China, pp. 125-128, 2000.

[6] H. Y. Gu and Y. Z. Zhou, "An HNM Based Scheme for Synthesizing Mandarin Syllable Signal", *Int. Journal of Computational Linguistics and Chinese Language Processing*, Vol. 13, No. 3, pp. 327-342, 2008.

[7] H. Y. Gu and C. Y. Wu, "Model Spectrum-progression with DTW and ANN for Speech Synthesis", in *Proc. ECTI-CON 2009*, Pattaya, Thailand, pp. 1010-1013, 2009.

[8] K. Tokuda, *et al.*, "Speech Parameter Generation Algorithms for HMM-based Speech Synthesis", *ICASSP*, Istanbul, Turkey, pp. 1315-1318, 2000.

[9] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based Approach to Multilingual Speech Synthesis", in Text to Speech Synthesis: New Paradigms and Advances, Editors: S. Narayanan and A. Alwan, Prentice Hall, NJ, pp. 135-153, 2004.

[10] Y. Qian, F. Soong, Y. Chen, and M. Chu, "An HMM-Based Mandarin Chinese Text-to-Speech System", *ISCSLP*, Singapore, pp. 223-232, 2006.

[11] Y. 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.

[12] O. Cappé and E. Moulines, "Regularization Techniques for Discrete Cepstrum Estimation", *IEEE Signal Processing Letters*, Vol. 3 (4), pp. 100-102, 1996.

[13] H. Y. Gu and S. F. Tsai, "A Discrete-cepstrum Based Spectrum-envelope Estimation Scheme and Its Example Application of Voice Transformation", to appear in *Int. Journal of Computational Linguistics and Chinese Language Processing*, 2009.

[14] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993