SPEECH SYNTHESIS USING ARTICULATORY-KNOWLEDGE BASED HMM STRUCTURE

HUNG-YAN GU, MING-YEN LAI, WEI-SIANG HONG

Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology, Taipei 106, Taiwan
E-MAIL: {guhy, M9615074, M10115035}@mail.ntust.edu.tw

Abstract:

In this paper, a different HMM structure is proposed to model the context-dependent spectral characteristics of a speech unit in order to improve synthetic speech fluency. Instead of using decision trees, we base on the articulatory knowledge of phonemes to reduce the huge amount of context combinations. To evaluate the proposed HMM structure, three Mandarin speech synthesis systems using different HMM structures are constructed for comparison. In these systems, prosodic parameters are generated with same ANN modules developed previously but spectral parameters are generated with HMMs of themselves. As to the synthesis of signal waveform, a same HNM (harmonic plus noise model) based synthesis module developed previously is used. According to the results of listening tests, the speech signal synthesized by using the proposed HMM structure is significantly more fluent than those synthesized by using other HMM structures. In addition, the average spectral distances measured between recorded and synthetic sentences show that the proposed HMM structure can indeed decrease the spectral distance as compared with other HMM structures.

Keywords:

Speech synthesis; HMM structure; Articulatory knowledge; Spectral fluency; Discrete cepstral coefficient; HNM

1. Introduction

Recently, HMM (hidden Markov model) have been adopted by many researchers to model the spectrum progression within a speech unit (e.g. phoneme or syllable) [1, 2, 3, 4]. To synthesize a sentence, the trained HMMs are first used to generate a spectral feature vector sequence. Then, speech signal is synthesized with the generated spectral vector sequence. Speech signals synthesized with HMMs usual have improved intelligibility and fluency. Furthermore, Tokuda, *et al*., have developed the programs, HTS, based on HTK for HMM based speech synthesis [2], and provide the source code of HTS for other researchers to download. Therefore, the efforts needed to study speech synthesis can be reduced a lot with HTS. However, one should know that speech signals synthesized using HTS without GV (global variance) matching are generally too smooth and are perceived as muffled [5].

In this study, we decide not to use HTS. Hence, we must develop programs to train HMMs and to generate spectral feature vector sequence. This is because we intend to develop a speech synthesis system that is flexible for adding extra functions. For example, one function is timbre transformation to transform the synthesized speech timbre from a female adult to a male child [6]. Another planned function is to synchronously play the synthesized speech signal with its corresponding phonetic symbols. This function is needed because we will install the developed speech synthesis system to a humanoid robot in the future. Besides the factor, adding extra functions, the pitch contours generated by HTS are not satisfactory for Mandarin speech synthesis according to our experience in using HTS and other researcher's study [7]. Therefore, a different method for generating pitch contours is adopted in our speech synthesis system.

In the previous study [4], we have once attempted to model the spectrum progression within a syllable with HMM. The synthesized speech signal is not fluent enough. Spectral discontinuities may be perceived at syllable boundaries. We think one reason is that the unit, syllable, is too large, and the quantity of different contextual dependencies between syllables is too large to be well modeled. Therefore, in this paper, we take smaller speech units, i.e. syllable initial (consonant) and syllable final (vowel, diphthong, or nasal ended vowel). Also, context-dependent HMMs are grouped and structured according to the phonetic symbol sequence labeled in the transcription files. The detail of the structuring method proposed here will be described in Section 2.

As a global view, the processing flow of our systems' synthesis stage is depicted in Figure 1. In Block (a), the input text is analyzed. In Block (b), ANN models are used to generate each syllable's pitch-contour parameters and duration value. The details are referred to [8]. In Block (c), the selection method will be explained in Section 2. In Block (d), the method proposed by Tokuda, *et al*., is adopted to

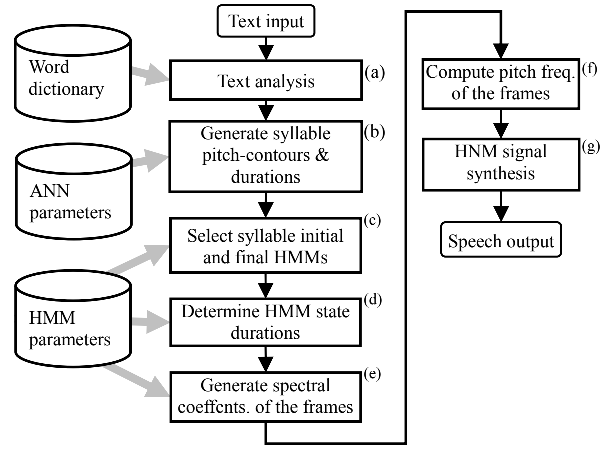
Text input
Word dictionary
Text analysis
(a)
ANN parameters
Generate syllable pitch-contours & durations
(b)
Select syllable initial and final HMMs
(c)
HMM parameters
Determine HMM state durations
(d)
Generate spectral coeffcnts. of the frames
(e)
Compute pitch freq. of the frames
(f)
HNM signal synthesis
(g)
Speech output

Figure 1. The main processing flow for the synthesis stage

compute each state's duration in frames [1]. In Block (e), the method, weighted linear interpolation, proposed in a previous study [4] is adopted. In Block (f), the pitch frequency of each voiced frame is computed in terms of the pitch-contour parameters. In Block (g), an HNM (harmonic plus noise model) based signal waveform synthesis method is adopted. Its details are referred to [6].

2. Context Classification and Combination

In speech recognition, the left-to-right structure as shown in Figure 2 is the basic HMM structure. Notice that this structure does not handle the contextual dependencies at the left and right boundaries. The performance (e.g. recognition rate) of such HMM structure would be degraded.

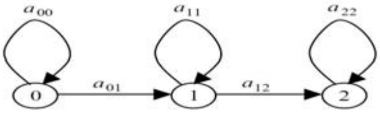

Figure 2. Left to right HMM structure

A mandarin syllable is conventionally divided into syllable initial (i.e. initial consonant) and final (i.e. final vowel cluster). Such dividing is helpful to decrease the huge number of left and right context combinations when syllable is adopted as the speech unit. However, the number of possible context combinations is still very large even when the speech units, syllable initial and final, are adopted.

At the left side of a syllable initial, the possibly encountered speech unit is the final of the last syllable or silence. Here, we classify the possible mouth-gestures at the ending of a syllable final into 11 classes. The details are as shown in Table 1 (including "sil", silence). By contrast, at the right side of a syllable initial, the speech unit encountered will be a syllable final. Here, we classify the possible mouth-gestures at the start of a syllable final into 8 classes. The details are as shown in Table 2. Because there are 21 different syllable initials in Mandarin, the number of possible context combinations for syllable initials are 12 × 21 × 8 = 2,016.

TABLE 1. ENDING-GESTURE CLASSIFICATION FOR SYLLABLE FINALS

Index	0	1	2	3	4	5	6	7	8	9	10	11
Gesture class	a	o	ə	e	i	u	yu	ii	er	n	ng	sil

TABLE 2. START-GESTURE CLASSIFICATION FOR SYLLABLE FINALS

Index	0	1	2	3	4	5	6	7
Gesture Classes	a	o	ə	e	i	u	yu	ii

Similarly, at the left side of a syllable final, the possible encountered speech unit is the final of the last syllable or silence if the current syllable has no initial consonant. In this case, the possible mouth-gestures at the ending of a syllable final are classified into 12 classes as shown in Table 1. In the other case, the speech unit at the left side may be an initial consonant. Here, we classify the possible consonants into 6 classes according their articulation positions. For example, the four consonants, b, p, m, f, are all articulated at the position, lip. Therefore, they are placed to the class, "b". Similarly, the four consonants, d, t, n, l, are all articulated at the position, alveolar ridge, and are therefore placed to the class, "d". In detail, the 6 consonant classes are as listed in Table 3.

TABLE 3. CLASSIFICATION OF SYLLABLE-INITIAL CONSONANTS

Index	0	1	2	3	4	5
Consonant Classes	b	d	z	zh	j	g

On the other hand, at the right side of a syllable final, the possible encountered speech unit may be silence or the final of the next syllable if the next syllable has no initial consonant. In this case, the possible mouth-gestures at the start of a syllable final are classified into 10 classes as shown in Table 1 without /n/ and /ng/. In the other case, the speech unit at the right side may be an initial consonant of the next syllable. Then, there would be 6 consonant classes as shown in Table 3. Since Mandarin has 37 different syllable finals, the number of context combinations for syllable finals is (12+6) × 37 × (10+6) = 10,656.

If we plan to build an HMM for each context combination, then the number of HMMs to be trained is 2,016 plus 10,656. This implies that we must record several times of 10,656 syllables when preparing the training sentences. Therefore, it is expensive to prepare such large amount of training sentences, and is thus impractical to adopt such context dependent HMMs. One solution proposed by others is to cluster the HMMs into a smaller number of clusters by using a decision tree, e.g. HTS. However, in this paper, we will study an HMM structuring method to solve the mentioned problem.

3. New HMM Structure

An original HMM structure, of six states, for a syllable final is dependent on both left and right contexts as drawn in Figure 3. The symbol, FY_XY, denotes an HMM for the syllable final, FY, that is just preceded by the context type, CX, and succeed by the context type, CZ. According to Tables 1 and 3, the number of context types for CX is 12 + 6 = 18. Similarly, according to Tables 1 (excluding /n/ and /ng/) and 3, the number of context types for CZ is 10 + 6 = 16. Since Mandarin has 37 different finals, FY, the number of context dependent HMMs for syllable finals is thus as large as 10,656 (18 × 37 × 16).

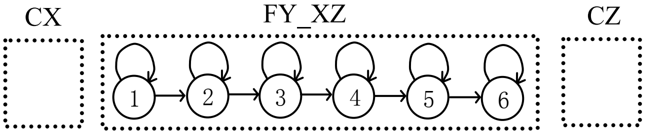

Figure 3. Left and right context dependent HMM structure

To decrease the cost and efforts needed for practical implementation (e.g. preparing large amount of training sentences), we hence study to restructure the HMM, FY_XZ, shown in Figure 3. The solution proposed here is to make the assumption that the front half of the six states (i.e. states 1, 2 and 3) in Figure 3 are dependent on the context type, CX, but not dependent on the context type, CZ. Similarly, we also assume that the back half, i.e. state 4, 5 and 6, are dependent on CZ but not dependent on CX. Accordingly, the context dependent HMM, FY_XZ, in Figure 3 can be decomposed into the two half (half context-dependent and size) HMMs, GY_X and HY_Z, shown in Figures 4 and 5, respectively.

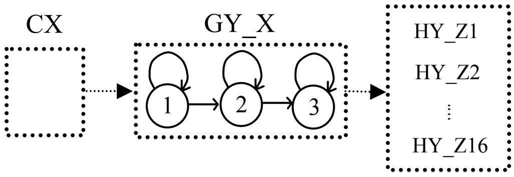

Figure 4. Left-context dependent HMM structure for the front half

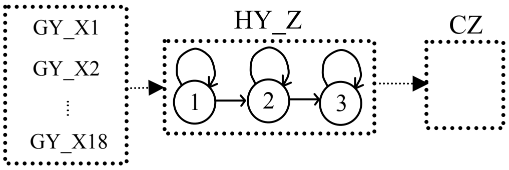

Figure 5. Right-context dependent HMM structure for the back half

Notice that the context preceding a syllable final is classified to 18 types. Therefore, we need 18 half HMMs, GY_X1, GY_X2, …, and GY_X18 (as shown in Figure 5), to model the front part of the syllable final, FY. Also, the context succeeding a syllable final is classified to 16 types. Therefore, we need 16 half HMMs, HY_Z1, HY_Z2, …, and HY_Z16 (as shown in Figure 4), to model the back part of the syllable final, FY. Consequently, the number of half HMMs required to model a syllable final is 18 + 16 = 34, and the total number of half HMMs needed to model 37 different syllable finals is 34 × 37 = 1,258. Apparently, 1,258 is much smaller than 10,656, the number of HMMs for modeling left and right context-dependent syllable finals.

For modeling the 21 syllable initials, similar assumptions made to syllable finals are also adopted here. Since the context preceding a syllable initial is classified to 12 types, 12 half HMMs are needed to model the front part of a syllable initial. Similarly, 8 half HMMs are needed to model the back part of a syllable initial. Consequently, the number of half HMMs required to model a syllable initial is 12 + 8 = 20, and the total number of half HMMs required to model 21 different syllable initials is 20 × 21 = 420. Apparently, 420 is much smaller than 2,016, the number of HMMs to model left and right context-dependent syllable initials.

4. Experimental Evaluation

4.1. Training stage

We invited a male adult to record 1,208 sentences in a soundproof room. The script is composed of randomly selected sentences from independent articles. Totally, there are 10,173 Mandarin syllables in these sentences. Here, the sampling rate adopted is 22,050 Hz. For labeling the recorded syllables, the package, HTK, was used first to perform forced alignment. Then, the software, WaveSurfer, is used to adjust syllable boundaries manually.

The signal file of each syllable is sliced into a sequence of frames. Frame width is set to 512 sample points and frame shift is 128 points. For each frame, a vector of 39 spectral parameters, i.e. discrete cepstral coefficients (DCC) [9], c_0, c_1, …, c_{38}, are extracted. The details for extracting DCC are referred to our previous work [10]. Additionally, the periodicity of a frame is also saved to an extra dimension, i.e. c_{39}. The value of c_{39} is set to 1 if the speech frame is periodic, and set to 0 if not periodic. In term of the dimension, c_{39}, whether an HMM state is voiced or unvoiced can be decided, and the voicing information is used in generating pitch contours. 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.

In this study, the front and back half-HMMs for syllable finals are all constructed with 3 states, and transited in the left to right manner as shown in Figures 4 and 5. Nevertheless, for syllable initials, only 2 states are used to construct each half-HMM. As to the number of Gaussian mixtures, just one mixture is placed on each state. To train the half HMMs, we have developed programs according to the algorithm of segmental K-means [11].

4.2. Speech synthesis processing

Our Mandarin speech synthesis system follows the processing flow shown in Fig. 6. In "text analysis" block, a sentence is read and parsed from the input each time. Then, the sentence is segmented into a sequence of words by looking up the word dictionary, and a phonetic syllable symbol is obtained for each character. Next, in the block "generate pitch-contours and durations", contextual data items are prepared for each syllable of the sentence first. Then, the contextual data are fed to two ANNs to generate a pitch-contour and a duration value for each syllable. The details about the ANN structure and input/output data items are referred to our previous work [8].

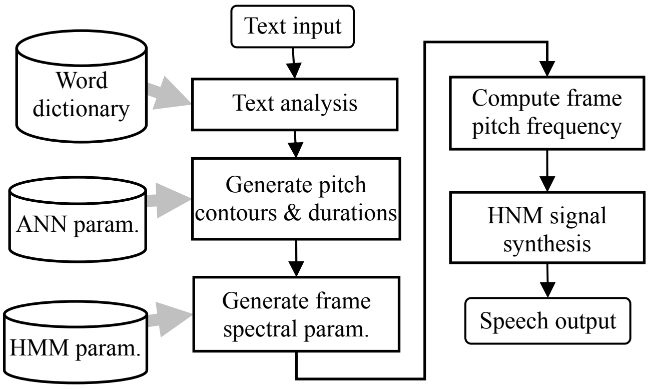
Text input
Word dictionary
Text analysis
ANN param.
Generate pitch contours & durations
HMM param.
Generate frame spectral param.
Compute frame pitch frequency
HNM signal synthesis
Speech output

Figure 6. Main flow of the Mandarin speech synthesis system

In the block, Generate frame spectral parameters, each syllable's pronunciation symbol is first split into its initial and final parts. Then, according to the id of a unit (initial or final) and its left and right context, two corresponding half HMMs can be retrieved from the trained HMM collection, and concatenated to form a full HMM for the unit. Next, the states of the HMM will be assigned some numbers of frames according to the 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 [12]. Now, consider how to generate each frame's spectral parameters, i.e. DCC. One commonly used method is based on MLE (Maximum likelihood Estimate) [12]. Here, we use a different generation method, called WLI (weighted-linear interpolation), which is proposed in a previous work [4].

In the block, Compute frame pitch frequency, each frame of a voiced unit (e.g. initial /m/ and final /a/) is assigned a pitch value. Here, the pitch-contour parameters generated by the ANN are Lagrange interpolated to compute a pitch value for each frame. Next, in the block, HNM signal synthesis, the DCC and pitch value generated for each frame of a unit are processed in frame order to synthesize speech signals. The details for signal synthesis are referred to a previous work [10].

4.3. Spectral distance measuring

The two speech synthesis systems constructed in this study are denoted as SYC and SYD. In the complete system, SYC, the left context of a front-half HMM (e.g. GY_X in Figure 4) and the right context of a back-half HMM (e.g. HY_Z in Figure 5) are both distinguished. That is, a corresponding half HMM (e.g. GY_X) is constructed for each different context type (e.g. CX). By contrast, in the downgraded system, SYD, the contexts preceding or succeeding a syllable are disregarded. That is, the left context of a front-half HMM for a syllable initial is not distinguished, i.e. only one front-half HMM is constructed for each syllable initial. Also, the right context of a back-half HMM for a syllable final is not distinguished, i.e. only one back-half HMM is constructed for each syllable final. In addition, for the purpose of comparison, the system constructed in our previous study [4] is denoted as SYP. In the system, SYP, each different Mandarin syllable is directly modeled with one or a few syllable-wide HMMs. The number of HMMs constructed for a syllable is dependent on the syllable's occurrence times in the training sentences. Here, the same training sentence set is used to train the HMMs for the three systems, SYC, SYD, and SYP.

In terms of the three systems, we experiment to measure spectral distances between corresponding frames of a recorded sentence and a synthetic sentence. In detail, label files of 50 test sentences are fed one after another to each system to obtain their corresponding synthetic speech files.

Then, each pair of recorded and synthetic speech files is analyzed to extract two sequences of DCC vectors. Next, every two DCC vectors from corresponding frames are taken to compute a geometric distance if both frames are detected to be voiced. In terms of the distances computed, a global average distance is then computed across 50 sentences' frames. As a result, we obtain the average spectral distances for the three systems as listed in Table 1. It is seen from Table 1 that the DCC vectors generated by the system, SYC, are closest to those analyzed from the recorded sentences whereas the DCC vectors generated by the system, SYP, are the farthest among the three systems. In addition, even only the contextual dependency between the initial and final of a syllable is modeled (i.e. System SYD), the measured spectral distance can still be improved a lot as compared with the system, SYP. These indicate that the half-HMM structures shown in Figures 4 and 5 can indeed help to better model the context dependent spectra at the boundary of two adjacent speech units.

TABLE 1. MEASURED AVERAGE SPECTRAL DISTANCES

System	SYC	SYD	SYP
Avg. dist.	0.633	0.640	0.732

4.4. Subjective listening tests

A short article not included to the training sentences is used here. This article is consisted of 70 syllables, and is fed to the three systems respectively to synthesize speech signal files. For convenience, the speech files synthesized by the three systems, SYC, SYD and SYP, are denoted as WC, WD and WP, respectively. These speech files can be accessed at the site, http://guhy.csie.ntust.edu.tw/hmmhalf/ .

In terms of the speech files, WC, WD and WP, listening tests are conducted to compare the fluency levels of these files. We invite 12 persons to participate in the tests. In the first run, the speech files, WC and WD, are played in random order to each of the participant. Similarly, WC and WP are played in the second run whereas WD and WP are played in the third run. In each run, each participant is requested to give a score to indicate which of the two files played is more fluent and preferred. The scores defined here are from -2 to 2. Among the 5 scores, 2 (-2) means the latter (former) played file is apparently more fluent than the former (latter) played file. The score, 1 (-1), means the latter (former) played file is slightly more fluent than the former (latter) played file, and 0 means the fluency level of the two files cannot be distinguished.

After listening tests, the scores given by the participants are reordered and averaged for the three runs respectively. The average scores obtained are as those listed in Table 2. From Table 2, it is seen that the average scores for the first and second runs are -0.833 and -0.417. We think these score values will become larger (much minus) if the participants are all familiar with the research field of speech synthesis. These minus scores indicate that the speech file WC is perceptually more fluent than the other two files, WD and WP. Therefore, the half HMM structures studied here is effective to synthesize more fluent speech signals. As to the average score, 0.250, obtained in the third run, its absolute value is the smallest, and may indicate that the difference in fluency level between WD and WP is not significant.

TABLE 2. AVERAGE SCORES OF THE LISTENING TESTS

Run	WC vs. WD	WC vs. WP	WD vs. WP
AVG	-0.833	-0.417	0.250
STD	0.718	0.900	0.866

5. Conclusions

In this paper, a different type of HMM structure, half (context dependent) HMM, is studied to model the context dependent spectral characteristics of a speech unit (syllable initial or final) in order to improve synthetic speech fluency. Instead of using decision trees to classify the huge amount of context combinations, we base on the articulatory knowledge of phonemes to decrease the number of context combinations.

The experiments conducted to evaluate the proposed HMM structure include spectral distance measuring and listening tests. The average spectral distances measured between recorded and synthetic sentences show that the half HMM structures can decrease the average distance from 0.732 to 0.633. In addition, according to the results of listening tests, the speech signals synthesized by using the half HMM structure are more fluent than those synthesized by using the other two types of HMM structures. Therefore, the half HMM structure is helpful to improve the synthetic speech fluency under the situation of insufficient training sentences.

Acknowledgements

This study is supported by National Science Council of Taiwan under the contract number, NSC 102-2221-E-011-129.

References

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM based speech synthesis", in Proc. Eurospeech, Budapest, Hungary, pp. 2347-2350, September 1999.

[2] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0", in Proc. 6th ISCA Workshop on Speech Synthesis, Bonn, Germany, pp. 294-299, August 2007.

[3] Z. J. Yan, Y. Qian, and F. K. Soong, "Rich context modeling for high quality HMM-based TTS", in Proc. INTERSPEECH, Brighton, UK, pp. 1755-1758, September 2009.

[4] H. Y. Gu, M. Y. Lai, and S. F. Tsai, "Combining HMM spectrum models and ANN prosody models for speech synthesis of syllable prominent languages", in Proc. ISCSLP, Tainan, Taiwan, Special Session 1, December. 2010.

[5] T. Toda and K. Tokuda, "Speech parameter generation algorithm considering global variance for HMM-based speech synthesis", in Proc. Eurospeech, Lisbon, Portugal, pp. 2801-2804, September 2005.

[6] H. Y. Gu and C. L. Tsai, "Integrating speaker -nonspecific timbre transformation to an HNM based speech synthesis scheme", Journal of the Chinese Institute of Engineers, Vol. 36, No. 3, pp. 371-381, 2013.

[7] C. C. Hsia, C. H. Wu, and J. Y. Wu, "Exploiting prosody hierarchy and dynamic features for pitch modeling and generation in HMM-based speech synthesis", IEEE trans. Audio, Speech, and Language Processing, Vol. 18, No. 8, pp. 1994-2003, 2010.

[8] 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, May 2009.

[9] O. Cappé and E. Moulines, "Regularization techniques for discrete cepstrum estimation", IEEE Signal Processing Letters, Vol. 3, No. 4, pp. 100-102, 1996.

[10] H. Y. Gu and S. F. Tsai, "A Discrete-cepstrum based spectrum envelope estimation scheme and its example application of voice transformation", Int. Journal of Computational Linguistics and Chinese Language Processing, Vol. 14, No. 4, pp. 363-382, 2009.

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

[12] 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.