A PITCH-CONTOUR GENERATION METHOD COMBINING ANN, GLOBAL VARIANCE, AND REAL-CONTOUR SELECTION

HUNG-YAN GU, KAI-WEI JIANG

Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan
E-MAIL: guhy@csie.ntust.edu.tw, m10015067@mail.ntust.edu.tw

Abstract:

Pitch contours are important for synthesizing highly natural speech signal. In this paper, we study a new pitch-contour generation method. The method proposed is to combine ANN prediction module with global-variance matching (GVM) and real contour selection (RCS) modules. Here, a syllable pitch contour is first analyzed and then transformed via discrete cosine transform (DCT) to a DCT-coefficient vector. Each sequence of DCT vectors analyzed from a training sentence plus contextual parameters are then used to train the ANN weights and GVM parameters. In pitch-contour generation experiments, we measure variance-ratio (VR) values for objective evaluations. The modules, GVM and RCS, are shown to be helpful to promote VR values. In addition, in subjective evaluation, the pitch-contour generation method, ANN + GVM, is shown to be more natural than the method, ANN only. Also, the method, ANN + GVM + RCS, is shown to be better than ANN + GVM.

Keywords:

Speech synthesis; pitch contour; discrete cosine transform; artificial neural network; global variance; contour selection

1. Introduction

The naturalness level of a synthetic speech signal is chiefly determined by the prosodic parameters, e.g. syllable durations, pitch contours, intensities. Among the prosodic parameters, pitch contours are especially important for obtaining higher naturalness level. Therefore, many methods have been proposed to generate the syllable pitch-contours for synthesizing a Mandarin Chinese sentence [1, 2, 3, 4, 5, 6]. Although HMM (hidden Markov model) is currently adopted by many researchers for studying speech synthesis [7, 8], the pitch contours generated by MSD-HMM (multi-space probability distribution HMM) are however not satisfactory enough as noticed in [3, 6].

In this paper, we propose a new method that combines three techniques, i.e. artificial neural network (ANN) [1, 2], global variance (GV), and real-contour selection (RCS), to generate syllable pitch-contours for synthesizing Mandarin sentences. It is intended that the naturalness level of synthetic speech can be further promoted by combining ANN with GV and RCS.

Historically, GV matching (GVM) is proposed by Toda and Tokuda [9] to adjust the HMM generated spectral coefficients in order to alleviate the phenomenon of spectral over-smoothing that lowers the synthetic-signal quality. Here, we find that the phenomenon of over-smoothing is also observable in ANN generated DCT (discrete cosine transform) coefficients that represent a pitch contour. Therefore, we think GVM may be helpful to promote the naturalness level of ANN generated pitch contours. In addition, we are motivated by the concept of target-speaker frame selection studied in [10] to improve the converted voice quality. Hence, we think it will be helpful to increase the naturalness level if an ANN generated and GV adjusted pitch contour, X, is further used to select a real pitch-contour (analyzed from an uttered syllable), Y, and then Y is used to replace X. To implement real-contour selection (RCS), a corresponding pool that collects real pitch-contours classified as of same context type as X must be prepared in the training stage.

As a global view, the processing flow for the training stage of our system is drawn in Figure 1. First, each syllable of each recorded sentence is analyzed to obtain its corresponding pitch contour. Then, each pitch contour is transformed to a DCT-coefficient vector of fixed dimensions. Next, the sequence of DCT vectors and their corresponding contextual parameters are used to train the ANN based pitch contour generation model. Besides training ANN, the sequence of DCT vectors are further analyzed to obtain the parameters needed for GVM. In addition, each pitch contour represented as a DCT vector is collected to different pools (called real-contour pools) according to its carrying syllable's context type.

On the other hand, the global view of the processing flow for pitch contour generation is drawn in Figure 2. First, a written sentence (text) is read in. Then, by looking up a dictionary, the pronunciation syllable and lexical tone of each

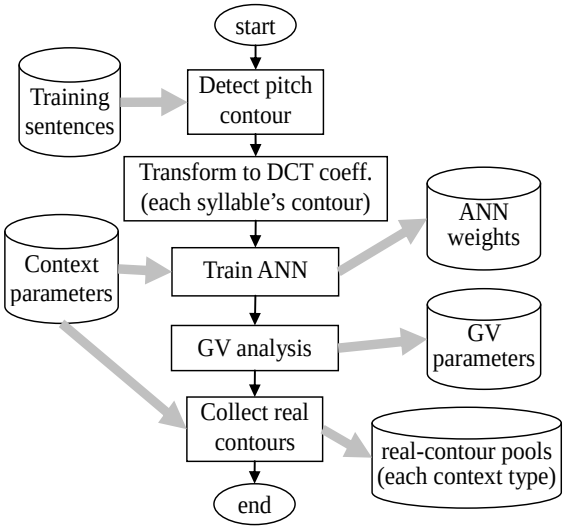

Figure 1. Main processing flow for the training stage

Chinese character is determined. In terms of the sequence of syllables and tones, contextual parameters are prepared for each syllable. Next, the contextual parameters for each syllable are feed to the ANN model to predict a pitch contour (i.e. DCT coefficients) for that syllable. In terms of the ANN predicted pitch contour, GV matching is performed with the saved GV parameters. In addition, in terms of the GV matched pitch contour, a nearest real pitch-contour is searched from the pool corresponding to that syllable's context type.

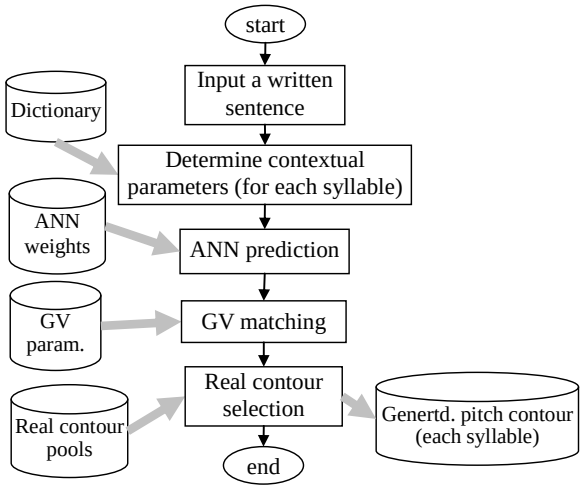

Figure 2. Main processing flow for the generation stage

2. Training of model parameters

As illustrated in Figure 1, the weights of the ANN model must be trained, the parameters for GV matching must be analyzed, and real pitch-contours (i.e. DCT vectors) must be collected to several pools for different context types.

2.1. Sentence recording and pitch contour detection

In this study, we invite a male speaker to utter 810 sentences in a soundproof room. The total number of syllables uttered in these sentences is 7,161. After recording, the sentences are first automatically labeled with the software package HTK. Then, the time boundaries of the syllables are manually checked and corrected with the software package WaveSurfer.

To detect the pitch contours of the syllables, we use the modules of SPTK included in HTS package [8]. The sampling rate adopted here is 22,050 Hz, and the frame shift is 110 sample points. After automatic detection, we find that many frames' pitch frequencies are erroneously detected. For example, the frequency detected for a voiced frame may be zero (i.e. decided to be unvoiced), half or double of the true frequency. Therefore, we have developed a pitch-contour tool program for semi-automatically or manually correcting the erroneously detected syllable pitch contours.

2.2. Discrete cosine transform

Notice that a syllable of an uttered sentence may be of length form 30 frames to around 80 frames. Hence, we decide to represent a syllable pitch-contour as a DCT coefficient vector with fixed dimensions. As to the number of dimensions, we select to use 24 dimensions. This decision is based on comparing some corresponding pairs of original and inversely transformed pitch contours with different dimensions.

In details, the formula adopted here to calculate DCT coefficients is the DCT-I type [11], i.e.

$$c(m) = x(0) + (-1)^m \cdot x(N-1) + 2 \cdot \sum_{k=1}^{N-2} x(k) \cdot \cos\left(\frac{m \cdot k \cdot \pi}{N-1}\right), \quad m = 0,1,...,23 \tag{1}$$

where $x(k)$ denotes the pitch frequency (in Hz) of the k-th frame, $c(m)$ denotes the m-th DCT coefficient, and N is the number of frames. Corresponding to formula (1), the formula for inverse DCT transform is

$$x(k) = \frac{1}{2(N-1)} \left[\begin{array}{l} c(0) + (-1)^k \cdot c(M-1) + \\ 2 \cdot \sum_{m=1}^{M-2} c(m) \cdot \cos\left(\frac{k \cdot m \cdot \pi}{M-1}\right) \end{array} \right], \quad k = 0,1,...,N-1 \tag{2}$$

where M denotes the number of DCT coefficients, i.e. M=24 here.

2.3. ANN training

The structure of the ANN designed here is illustrated in Figure 3, i.e. a recurrent neural network. The input layer has 28 nodes to input 8 contextual parameters, and the output layer has 24 nodes to output 24 DCT coefficients representing a syllable pitch contour. The 8 contextual parameters include: (a) tone and syllable-final class of previous syllable; (b) tone, syllable initial, and syllable final of current syllable; (c) tone and syllable-initial class of next syllable; (d) time-progress index. The details for the classification of syllable initials and finals are referred to our previous work [12]. In addition, the number of nodes to be placed in the hidden layer must be decided. We had tested it from 12 to 20 nodes with the first 750 recorded sentences, and find that 16 nodes is the best choice according to our experiment results.

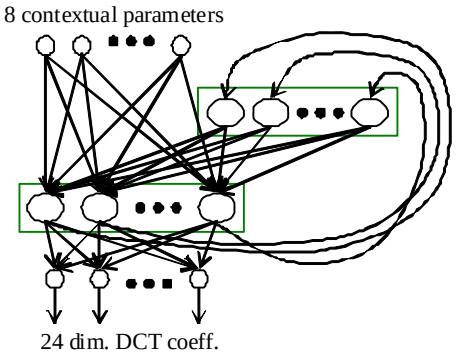

Figure 3. The structure of the ANN designed here

2.4. Analysis of GV parameters

GV matching is originally used to adjust the spectral coefficients of a sequence of speech frames [9]. Here, GVM is however performed in the speech unit, syllable, instead of frame. This is because the pitch contour of a syllable is only represented as one DCT vector of 24 dimensions. Suppose that the length of a sentence is from 4 syllables to 20 syllables. Then, only 4 to 20 DCT vectors are used to compute each dimension's variance value for a sentence. To estimate the variance of the *i*-th dimension for the *k*-th sentence, the formula is

$$v_i^k = \left[\sum_{j=1}^{n(k)} (c_i^k(j) - m_i^k)^2 \right] \Big/ n(k), \quad (3)$$

where $n(k)$ denotes the number of syllables in the *k*-th sentence, $c_i^k(j)$ denotes the DCT coefficient of the *i*-th dimension for the *j*-th syllable pitch-contour, and m_i^k denotes the mean value of $c_i^k(j)$, j=1, …, $n(k)$.

Then, the global variance for the *i*-th dimension across the 750 training sentences is estimated with the formula

$$g_i = \frac{1}{N} \sum_{k=1}^{N} v_i^k , \quad (4)$$

where N denotes the number of training sentences (i.e. 750 here), and g_i is the estimated global variance for the *i*-th dimension.

2.5. Collection of real pitch-contours

To implement real contour selection, we must prepare a real pitch-contour pool for each type of context combination in the training stage. How to define context types? First, pitch declining in sentence intonation is a well known phenomenon. Hence, we decide to divide the sequence of syllables of each training sentence into three segments. The syllables in the leading segment would have higher pitches whereas the syllables in the tail segment would have lower pitches.

Secondly, we think one of the chief factors that influence the height and shape of a syllable's pitch contour is the tone combinations of previous, current and next syllables. For example, let the (*j*-1)-th syllable of a sentence is uttered in tone P_{j-1}, the *j*-th syllable is uttered in tone P_j, and the (*j*+1)-th syllable is uttered in tone P_{j+1}. Then, the tone combination index for the *j*-th syllable is calculated as $25 \times P_{j-1} + 5 \times P_j + P_{j+1}$. Totally, there are 125 tone combination types since a syllable may be uttered in one of the five tones in Mandarin. If j=1, i.e. the first syllable, P_{j-1} is defined to be the neutral tone here. Similarly, if the *j*-th syllable is the last syllable, P_{j+1} is also defined to be the neutral tone.

Considering the two factors mentioned above, we define 3×125 = 375 context types for real pitch-contour classification. Therefore, we set up 375 pools to collect the real pitch-contour DCT vectors. By using the 750 training sentences, we throw each syllable's pitch-contour DCT vector to one of the 375 pools according to that syllable's context type.

3. Pitch contour generation and experimental evaluations

3.1. Pitch contour generation

According to Figure 2, a sequence of syllables and lexical tones are determined first for an inputted Chinese sentence. Then, the module, ANN prediction, is used to predict 24 DCT coefficients representing a pitch contour for each syllable whose contextual parameters are fed in. After

each syllable's pitch contour is predicted, the pitch-contour DCT vectors for the sequence of syllables are then adjusted in the module, GV matching. For each syllable, the formula used to match global variance is

$$\hat{c}_i = (c_i - m_i)\left[(w \cdot \sqrt{g_i / v_i}) + 1\right] + m_i \ , i = 0,1,...,23, \quad (5)$$

where c_i denotes the i-th dimension of an ANN-predicted DCT vector, m_i and v_i denote the mean value and variance, respectively, for those c_i across the sentence's syllables, g_i is as estimated in Formula (4), and w is the matching-strength weight whose value is set between 0 to 1.

After GVM, the module, real contour selection, is executed next. Let X_j denotes the GV-adjusted DCT vector for the j-th syllable of the sentence. For X_j, its corresponding context type, m_j, is determined first according to the segment it locates and the tone combination index of its adjacent three syllables as described in Section 2.5. Then, the real pitch-contours in the pool numbered m_j, are fully searched to find a DCT vector Y_j who is nearest to X_j in terms of a geometrical distance measure. Then, Y_j is used to replace X_j.

It is interesting to study the effects of the two blocks, "GV matching" and "Real contour selection" in Figure 2. Therefore, we experiment pitch contours generation here with six different methods that use or not use the two blocks mentioned, and use different weight values, w, in Formula (5). The symbols, MA, MB, MC, MD, ME and MF are coined here to denote the six methods. In details,

MA: ANN prediction but no GVM and RCS;
MB: ANN and GVM with w=0.33 but no RCS;
MC: ANN and GVM with w=0.5 but no RCS;
MD: ANN and RCS but no GVM;
ME: ANN, GVM with w=0.33, and RCS;
MF: ANN, GVM with w=0.5, and RCS.

3.2. Objective evaluations

For inside tests, the 750 recorded sentences used to train the ANN model and GV parameters are still used here to measure average geometric distances (between original and generated DCT vectors) and variance ratios. For outside tests, only the remaining 60 sentences that are not used in the training stage are used for measuring. The first result is that the measured average geometric distances do not have significant differences among the six methods. Therefore, the measure, variance ratio (VR), previously proposed for comparing converted-voice quality [13], is adopted here to compare the six methods. The formula to calculate VR is

$$VR = \frac{1}{L}\sum_{k=1}^{L}\frac{1}{D}\cdot\sum_{d=1}^{D}\frac{\hat{\sigma}_k^d}{\sigma_k^d} \ , \quad (6)$$

where L denotes the number of syllable-final classes in Mandarin (here L =36), D denotes the number of dimensions in a DCT vector, $\hat{\sigma}_k^d$ denotes the variance calculated from the d-th dimension of the generated pitch-contour DCT vectors that superimpose the k-th syllable-final class, and σ_k^d denotes the variance calculated from the d-th dimension of the analyzed (from recorded sentences) DCT vectors that superimpose the k-th syllable-final class. Notice that D is 23 here because the coefficient, c_0, in an ANN generated DCT vector is not modified by GVM and not replaced by RCS.

The measure VR values for the six methods are depicted in Figure 4. According to these VR values, it can be found that the pitch-contour DCT vectors generated by ANN (method MA) indeed have very low VR values just around 0.1. The VR values would be significantly increased if GVM (method MB or MC) or RCS (method MD) is applied to the ANN generated DCT vectors. In addition, the VR values become further higher if GVM and RCS are both applied in cascading (method ME or MF). Notice that the trend just mentioned is consistently seen in the two curves of Figure 4 for both experiments using inside and outside sentences. Therefore, GVM and RCS are very effective to improve the phenomenon of over-smoothed DCT coefficients.

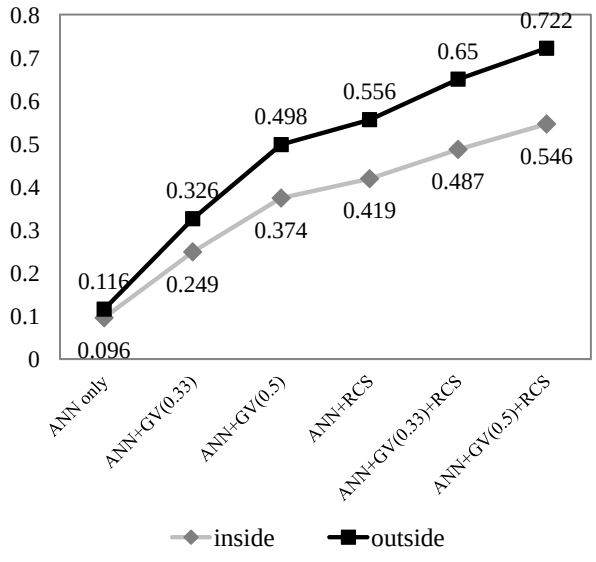

Figure 4. VR values measured under different generation methods

3.3. Subjective evaluations

To achieve higher naturalness level, a generation method should generate pitch-contours that are perceived as having correct tones and sufficient pitch range and curvature. Here, we evaluate the six generation methods by listening tests.

Two groups of persons are invited to participate in the listening tests. The first group consists of 11 persons who have experience in speech signal processing. In contrast, the second group consists of 11 persons too but they have no experience in speech signal processing.

For preparing synthetic speech files, three short articles are randomly selected, and each of the six generation methods (from MA to MF) is used to generate the syllable pitch-contours for each article, respectively. Then, the generated pitch contours by the six methods and the other prosodic parameters (syllable durations and intensities) are fed to the signal synthesis module to synthesize 6 speech signal files for each article [14]. In addition, we have converted the generated pitch contours from a male's pitch to a female's pitch by using a method commonly adopted in voice conversion [10, 13]. Then, the converted pitch contours and previously trained spectral HMM using this female's uttered sentences are used to synthesize that female's speech signal files. Hence, each of the six generation methods has 6 (3 articles × 2 speakers) speech files synthesized.

Here, listening tests are executed by requesting each participant to compare two played synthetic speech files and then give a score to indicate which is more natural. If the former (latter) is apparently natural than the latter (former), the score, 1 (5), should be given. If the former (latter) is slightly natural than the latter (former), the score, 2 (4), should be given. Otherwise, the score, 3, is given to indicate that the two played speech files cannot be distinguished in naturalness level.

Notice that the combination number for taking any two methods from the six generation methods is 15, which require too many efforts to execute listing tests. Hence, we select only five method pairs for listening tests, i.e. (a) MA vs. MB, (b) MB vs. MC, (c) MA vs. MC, (d) MB vs. ME, and (e) MC vs. MF. For each method-pair, each of the participants would listen to 6 synthetic speech file pairs in a sequence, and give a score for each speech file pair. After listening tests, the scores that are collected from comparing a same pair of speech files and given by the participants from a same group are averaged. Then, we consider the average score as a voting, i.e. we add one vote to the former generation method if the average score is less than 3, and add one vote to the latter method if the average score is greater than 3. Since there are 6 speech files synthesized for each of the six generation methods and the scores given by the two participant groups are averaged separately, the total number of votes for the comparison of each method pair is 12. As our experiment results, the votes obtained by the two methods of each method pair are depicted in Figure 5.

According to the voting results shown in Figure 5, it can be seen that the votes for MA vs. MB are 2 vs. 10, the votes for MB vs. ME are 3 vs. 9, and the votes for MC vs. MF are 4 vs. 8. Therefore, the method MB (ANN and GVM) will generates more natural pitch contours than MA (ANN only). In addition, the module, RCS, is shown to be effective in raising the naturalness level according to the voting results of MB vs. ME and MC vs. MF. On the other hand, the votes for MB vs. MC are 6 vs. 6 and the votes for MA vs. MC are 7 vs. 5. Therefore, there is no significant difference in naturalness level between MB and MC that only differ in GVM weigh values.

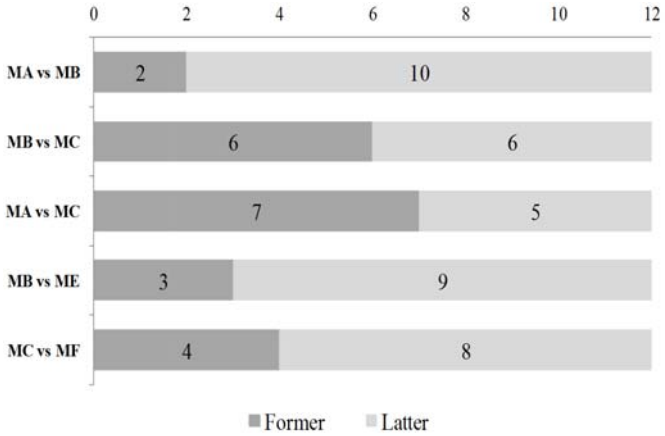

Figure 5. Voting results for the comparisons of the 5 method pairs

4. Concluding remarks

We find that the phenomenon of over-smoothing exists in the ANN generated DCT coefficients that representing a pitch contour. Therefore, in this paper, we attempt to promote the naturalness level of ANN generated pitch contours by cascading two more processing modules, i.e. GVM and RCS, to the ANN prediction module.

In objective evaluation, VR is used to measure the level of over-smoothing. According to the measured VR values, it is found that both modules, GVM or RCS, are helpful to raise VR values significantly. Hence, GVM and RCS can indeed help to alleviate the problem of over-smoothed DCT coefficients. Moreover, the VR value will be further raised if both GVM and RCS are cascaded.

In subjective evaluation, we select five pairs of pitch contour generation methods to compare their naturalness level. After listening tests, the scores given by the participants are averaged respectively for different speech file pair and participant groups. Then, each average score is considered as a voting of naturalness level. Consequently, we find that the method MB (ANN and GVM) is voted to be better than MA (ANN only). In addition, the method ME is voted to be better than MB, and the method MF is voted to be better than MC. That is, RCS as used in ME and MF is indeed effective for raising the naturalness level.

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. 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, pp. 309-324, 2004.

[3] 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.

[4] H. Y. Gu and C. C. Yang, “An HMM based pitch-contour generation method for Mandarin speech synthesis”, Journal of Information Science and Engineering, Vol. 27, No. 5, pp. 1561-1580, 2011.

[5] M. Dong, K. T. Lua, “Pitch contour model for Chinese text-to-speech using CART and statistical model,” Int. Conf. on Spoken Language Processing, Denver, USA, pp. 2405-2408, 2002.

[6] L. Gao, Z. H. Ling, L. H. Chen, and L. R. Dai, “Improving F0 prediction using bidirectional associative memories and syllable-level F0 features for HMM-based Mandarin speech synthesis”, Proceeding of ISCSLP, Singapore, pp. 275-279, 2014.

[7] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Simultaneous modeling of spectrum, pitch and duration in HMM based speech synthesis”, Proceeding of EUROSPEECH, Budapest, Hungary, pp. 2347-2350, 1999.

[8] 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”, Proceeding of 6-th ISCA Workshop on Speech Synthesis, Bonn, Germany, pp. 294-299, 2007.

[9] T. Toda and K. Tokuda, “A speech parameter generation algorithm considering global variance for HMM-based speech synthesis”, IEICE trans. INF. & SYST., VOL. E90–D, NO.5, May 2007.

[10] T. Dutoit, A. Holzapfel, M. Jottrand, A. Moinet, J. Perez, and Y. Stylianou, “Towards a voice conversion system based on frame selection”, Int. Conf. Acoustics, Speech, and signal Processing, Honolulu, Hawaii, pp. 513-516, 2007.

[11] A. V. Oppenheim and R. W. Schafter, Descrete-time Signal Processing, second ed., Prentice-Hall, 1999.

[12] H. Y. Gu, Y. Z. Zhou, and H. L. Liau, “A system framework for integrated synthesis of Mandarin, Min-nan, and Hakka speech”, Int. Journal of Computational Linguistics and Chinese Language Processing, Vol. 12, No. 4, pp. 371-390, 2007.

[13] E. Godoy, O. Rosec, and T. Chonavel, “Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora”, IEEE trans. Audio, Speech, and Language Processing, Vol. 20, No. 4, pp. 1313-1323, 2012.

[14] H. Y. Gu, M. Y. Lai, and W. S. Hong, “Speech synthesis using articulatory-knowledge based HMM structure”, Int. Conf. on Machine Learning and Cybernetics, Lanzhou, China, pp. 371-376, 2014.