IMPROVED TIME-FREQUENCY TRAJECTORY EXCITATION MODELING FOR A STATISTICAL PARAMETRIC SPEECH SYNTHESIS SYSTEM

Eunwoo Song, Young-Sun Joo, and Hong-Goo Kang

Department of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea

ABSTRACT

This paper proposes an improved time-frequency trajectory excitation (TFTE) modeling method for a statistical parametric speech synthesis system. The proposed approach overcomes the dimensional variation problem of the training process caused by the inherent nature of the pitch-dependent analysis paradigm. By reducing the redundancies of the parameters using predicted average block coefficients (PABC), the proposed algorithm efficiently models excitation, even if its dimension is varied. Objective and subjective test results verify that the proposed algorithm provides not only robustness to the training process but also naturalness to the synthesized speech.

Index Terms— Statistical parametric speech synthesis, time-frequency trajectory excitation (TFTE), slowly evolving waveform (SEW), predicted average block coefficient (PABC)

1. INTRODUCTION

HMM-based statistical parametric speech synthesis systems have been successfully deployed in many applications because of their reasonable performance, even with a small database. However, the perceptual quality of synthesized speech is still unsatisfactory, mainly due to limitations in vocoding, the accuracies of acoustic models, and over-smoothed output [1]. Although a deep neural network (DNN) was introduced to replace HMM to enhance the accuracies of acoustic models and to relieve the over-smoothing problem [2-5], it is still unclear to understand the impact of vocoding techniques when they are combined with statistical models. In other words, how to design speech analysis/synthesis algorithm and how to parameterize the speech parameters suitably for the HMM/DNN-based training process need to be considered carefully.

The synthesized quality of a pulse or noise (PoN) model-based TTS system is buzzy and unnatural because the excitation signal is only modeled by either pulse or noise components [6]. To reduce the buzziness problem, various ways of mixed excitation models have been adopted [7][8]. By separating the whole frequency band into several fixed sub-bands, the excitation signal of each sub-band is represented by either PoN or band aperiodicities (BAP) [9][10]. However, it cannot fully represent the time-varying periodicity of various types of phonetic information, which makes the perceptual quality still buzzy or noisy.

To improve the naturalness of synthesized speech, a pitch-dependent time-frequency trajectory excitation (TFTE)-based TTS system was proposed [11][12]. The TFTE has an advantage in that it can represent the time-varying characteristics of phonetic information by decomposing it into slowly evolving waveform (SEW) and rapidly evolving waveform (REW) [13]. The SEW, the most important parameter in the TFTE-based scheme, represents the slowly varying components of excitation, such as voiced portion. On the other hand, the REW represents the remaining noisy components. Although utilizing the SEW/REW decomposition makes it more efficient to suitably estimate the periodicity in a unit of an individual frequency bin, it still has a problem of dimensional variation in the training process. Note that the number of parameters to be modeled in each pitch epoch is varied because of the pitch-dependent analysis paradigm. To solve the problem, only a fixed number of SEW magnitudes, typically in a low-frequency band, as well as the polynomial coefficients of the REW magnitude are used for the training process [12]. In the synthesis step, the remaining SEW magnitude for a high-frequency band is reconstructed by subtracting the REW magnitude from the normalized excitation. Therefore, the TTS system cannot fully utilize the maximum advantages that could be achieved by introducing the SEW/REW decomposition.

By adopting the predicted average block coefficient (PABC) technique [14], this paper proposes an improved parameterization method, which can appropriately regenerate the TFTE regardless of the time-varying feature dimensions. To make the process of training the TFTE parameters unaffected by the dimensionality problem, the average component of the SEW magnitude is first predicted by the SEW from the previous frame. As the temporal correlation of the SEW is very high, the redundancy of the SEW magnitude can be significantly reduced. The PABC-SEW is then obtained by subdividing the predicted average (PA) magnitude into a fixed number of frequency bands, and performing the discrete cosine transform (DCT) in each PA sub-band. As the DCT has good decorrelation and energy compactness properties, most information related to the PA magnitude is concentrated within the first few PABCs. Furthermore, the analysis of the remaining PABCs shows that they follow the normal distribution, which can be easily modeled by Gaussian random variables. As a result, training can be successfully accomplished by the fixed number of PABC-SEWs and REWs. Experimental results also verify that the proposed algorithm provides both robustness to the training process and naturalness to synthesized speech compared to the conventional algorithm.

This paper is organized as follows. Section 2 describes the TFTE algorithm with the conventional modeling method. Section 3 describes the proposed PABC-TFTE modeling method and verifies its advantages. Section 4 presents the experiments and results obtained by performing objective and subjective tests, and the conclusions are presented in the final section.

The authors would like to thank Microsoft Research Asia for funding this project through the Ministry of Knowledge Economy of South Korea.

2. SPEECH SYNTHESIS USING TIME-FREQUENCY TRAJECTORY EXCITATION

2.1. Time-Frequency Trajectory Excitation

TFTE uses a time-frequency surface to represent the voicing characteristics of an evolving excitation waveform. It extracts the periodicity of each individual frequency bin by decomposing a single pitch-based excitation signal into slowly and rapidly varying components.

Let $u(n, \phi)$ denote a periodic function with ϕ extracted at the n-th frame, then the TFTE signal can be represented as follows:

$$u(n,\phi) = \sum_{k=1}^{P(n)/2} [A_k(n)\cos(k\phi) + B_k(n)\sin(k\phi)], \quad (1)$$

where a phase ϕ is defined as $\phi(m) = 2\pi m/P(n)$ with a pitch period $P(n)$, and $A_k(n)$ and $B_k(n)$ are the k-th discrete time Fourier series (DTFS) coefficients of the excitation signal [13].

In every frequency bin, the periodic signal $u(n, \phi)$ is further decomposed into SEW and REW by applying a low-pass filter to the time-domain axis. The SEW component is obtained as follows:

$$u_{SEW}(n,\phi) = \sum_{m=1}^{M} h(m)u(n-m,\phi), \quad (2)$$

where $h(m)$ is the M-th order low-pass filter. Using the orthogonality, the REW is obtained by subtracting $u_{SEW}(n, \phi)$ from $u(n, \phi)$ as:

$$u_{REW}(n,\phi) = u(n,\phi) - u_{SEW}(n,\phi). \quad (3)$$

Therefore, it can be concluded that the SEW and REW represent the periodic and remaining noisy components of TFTE in each frequency bin, respectively.

2.2. TFTE Modeling for Speech Synthesis

To apply the TFTE model to the statistical parametric speech synthesis system, the SEW/REW coefficients should be adjusted to have fixed dimensions. Note again that the number of SEW/REW coefficients vary depending on the pitch period.

In the previous work given in [12], an approach that had been utilized in a low bit-rate speech coding technique was introduced. In the analysis step, only a small number of the low-frequency SEW magnitudes were used for representing voiced components. The REW magnitude was parameterized by the coefficients of the Legendre orthonormal polynomials because it was well-known that a power contour model was good enough to represent noisy components in perceptual aspects [13]. In the synthesis step, the polynomial coefficients were transformed into the REW magnitude. The high-frequency SEW magnitudes were reconstructed by subtracting the REW magnitude from one, and they were then combined to the modified low-frequency SEW magnitudes to recover the full frequency band information of SEW. Fig. 1-(b) shows an example of the reconstructed SEW magnitude using the above approach.

As we can guess from the figure, the synthesized quality of the method is unsatisfactory. It often creates buzzy sound, primarily because of the inaccurate periodicity. As the spectral trajectory in high-frequency region is too smooth, the synthesized speech is often unintelligible. It is more noticeable if the contents require a large amount of high-frequency components. Therefore, an efficient modeling algorithm that can appropriately represent the SEW trajectory should be designed to improve the naturalness of synthesized speech.

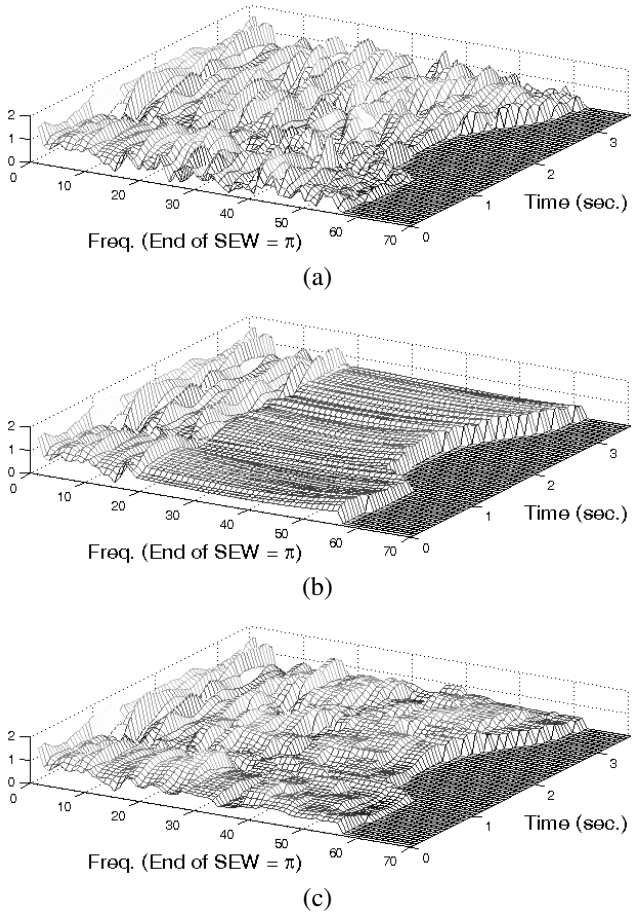

Fig. 1. SEW magnitude: original (a), as well as, reconstructed by the conventional (b) and proposed (c) algorithms.

3. IMPROVED TFTE MODELING FOR A STATISTICAL PARAMETRIC SPEECH SYNTHESIS SYSTEM

This section describes an improved TFTE modeling technique that is appropriate for the statistical parametric speech synthesis system. The advantage of the proposed algorithm is verified by comparing the trainability of the TFTE parameters to the conventional algorithm.

3.1. PABC-TFTE Modeling for Speech Synthesis

To improve the efficiency of the TFTE training process, it is very important to reduce the redundancy of the TFTE parameters. As the SEW magnitude is highly correlated with the adjacent SEW frames, the redundancy of the SEW magnitude can be significantly reduced by predicting the average magnitude component from the previous

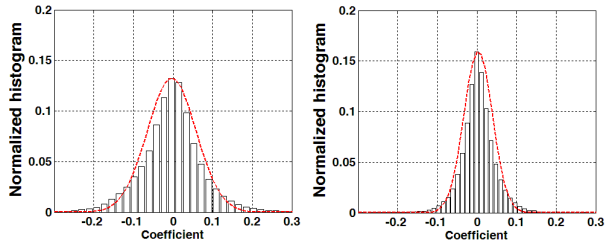

Fig. 2. Normalized histogram of the remaining PABCs in the 2-nd (left) and the 18-th (right) PA sub-block.

frame. The PA at the ϕ-th frequency bin is obtained as follows:

$$u_{PA}(n,\phi) = u_{SEW}(n,\phi) - u_{res}(n-1,\phi), \tag{4}$$

$$u_{res}(n-1,\phi) = \hat{u}_{SEW}(n-1,\phi) - u_{ave}(n-1), \tag{5}$$

where $\widehat{u}_{SEW}(n-1,\phi)$ represents an interpolated SEW magnitude at the ϕ-th frequency bin in the previous frame, and $u_{ave}(n-1)$ is the average SEW magnitude of the previous frame. As the difference between $u_{SEW}(n,\phi)$ and $u_{res}(n-1,\phi)$ becomes an average magnitude component of the current frame, it represents the smoothed version of the SEW magnitude. This process can be helpful to improve the efficiency of applying DCT, as discussed below.

The PA magnitude is then divided into K number of frequency sub-blocks:

$$\begin{bmatrix} c_{k,1} \\ \vdots \\ c_{k,J_k} \end{bmatrix}^T = \begin{bmatrix} u_{PA}(n, J_{k-1}+1) \\ \vdots \\ u_{PA}(n, J_{k-1}+J_k) \end{bmatrix}^T, \; 1 \le k \le K, \tag{6}$$

where $c_{k,j}$ denotes the j-th PA magnitude of the k-th sub-block, and J_k denotes a length of the k-th sub-block that satisfies the following condition:

$$\sum_{k=1}^{K} J_k = {}^{P(n)}/_{2}, \tag{7}$$

where $P(n)/_2$ is the length of the SEW. Each PA sub-block is then transformed with the DCT:

$$C_{k,m} = \frac{1}{J_k} \sum_{j=1}^{J_k} c_{k,j} \cos\left(\frac{\pi}{J_k}(j-0.5)(m-1)\right), \quad 1 \le m \le J_k \tag{8}$$

where $C_{k,m}$ is defined as the PR block coefficient (PABC), which represents the m-th DCT coefficient of the k-th sub-block.

As the DCT has good decorrelation and energy compactness properties [14], most information is concentrated within the first few PABCs. On the other hand, the remaining coefficients that have normal distributions are less important. Fig. 2 depicts examples of the distribution of the remaining PABCs. Therefore, the coefficients can be easily modeled by a single Gaussian. In summary, the full frequency band SEW magnitude can be trained with only several PABCs, but the remaining coefficients are generated by Gaussian random variables in the synthesis step. Fig. 1-(c) is an example of the reconstructed SEW magnitude that only uses the first coefficient of each PA sub-block and the generated random variables. Compared with the conventional approach (b), it is clear that the proposed algorithm recovers the SEW magnitude very well.

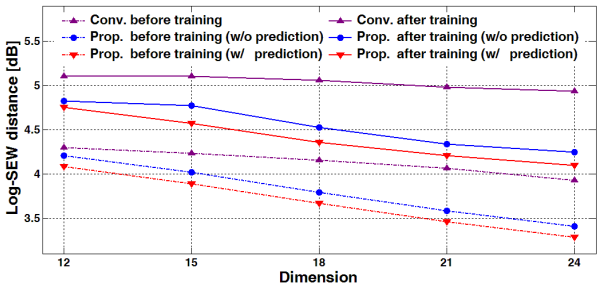

Fig. 3. Log-SEW distance before/after training.

3.2. Advantages of PABC-TFTE

This section describes the advantages of utilizing the PABC-TFTE for a statistical parametric speech synthesis system. To evaluate the effectiveness of the proposed algorithm in comparison with the conventional one, we measure the log-SEW distance between the original and reconstructed SEW magnitudes. The log-SEW distance is defined as:

$$D_{SEW} = \frac{1}{N}\sum_{n=1}^{N}\left\{\sqrt{\frac{1}{M_n}\sum_{k=1}^{M_n}\left(l_{ori}(n,k) - l_{syn}(n,k)\right)^2}\right\}, \tag{9}$$

where N denotes the number of frames and M_n represents the length of the SEW at the n-th frame; $l_{ori}(n,k)$ and $l_{syn}(n,k)$ represent the original and reconstructed log-SEW magnitudes (dB) in each frame and frequency bin, respectively. Note that the dynamic time warping technique is used to compensate for the durational mismatch between the original and reconstructed signals [15].

Fig. 3 represents the log-SEW distance depending on the SEW dimension before and after training. The results of both cases verify the effectiveness of the proposed algorithm in three ways. First, applying the prediction to the SEW magnitude has a merit in the modeling aspect, which reduces the reconstruction error of the SEW. Second, the results show that the reconstructed SEW of the proposed algorithm has much smaller error than that of the conventional one. Furthermore, the error significantly decreases in the proposed case when the SEW dimension becomes higher. Last, the amount of error increments before and after training is large in the conventional case due to the disadvantage of trainability, which is discussed in section 4.2. From the results, we confirm that the proposed PABC-TFTE not only reduces the error during reconstructing the excitation signal, but it also provides robustness to the training for excitation parameters.

4. EXPERIMENTS

4.1. Experimental Setups

To evaluate the effectiveness of the proposed algorithm, we constructed a context-dependent HMM-based Korean TTS system [16]. In total, 2,950 utterances recorded by Korean male speaker were used for training. The speech signals were sampled at 16 kHz, and each sample was quantized by 16 bits. A grapheme-to-phoneme (G2P) converter was also applied by the Korean standard pronunciation grammar and the context information-labeling program. More setup details are given in [12]. In the objective and subjective tests, twenty utterances not included in the training sets were used.

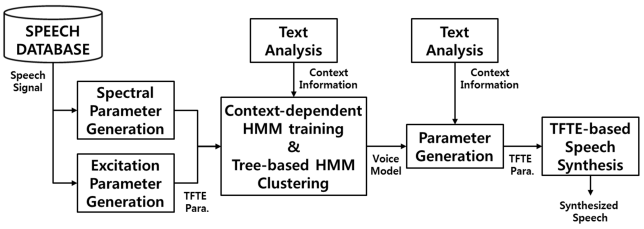

Fig. 4. Block diagram of the TTS system using TFTE.

Fig. 4 depicts a block diagram of the whole system. In the analysis step, the frame length is set to 20ms, and the spectral and excitation parameters are extracted every 5 ms. The 24-dimension line spectral frequencies (LSFs) are extracted for the spectral parameter. On the other hand, the 18-dimension PABC-SEWs and 4-dimension REW polynomial coefficients are extracted for the excitation parameters. The fundamental frequency (F0) and energy are also extracted for the HMM training. Table 1 summarizes the dimensions of each parameter. In the synthesis step, all parameters are generated by the context-dependent HMMs. The generated PABC-SEWs compose the SEW magnitude with Gaussian random variables. On the other hand, the REW magnitude is recovered by the generated polynomial coefficients. The phase extracted from speech is used for the SEW phase; on the contrary, the REW phase is randomly selected. The TFTE is then reconstructed from the SEW and REW with its pitch period. Finally, the single pitch-based speech signal is synthesized by the generated LSFs and TFTE.

4.2. Objective Test Result

To evaluate performance quantitatively, we measured the trainability of the proposed PABC-TFTE compared to the conventional one. It is defined by the normalized mean square error (NMSE) between the excitation parameters obtained from the original speech and those generated from the trained HMMs:

$$NMSE = \frac{1}{N}\sum_{n=1}^{N}\sqrt{\frac{\sum_{k=1}^{K}\left(x_{ori}(n,k)-x_{gen}(n,k)\right)^2}{\sum_{k=1}^{K}\left(x_{ori}(n,k)\right)^2}}, \quad (10)$$

where N and K denote the number of frames and the dimensions of the parameter, respectively; $x_{ori}(n,k)$ and $x_{gen}(n,k)$ denote the excitation parameters extracted from the original speech and generated by the trained HMMs, respectively.

Fig. 5 represents the average NMSE with a 95% confidence interval for each excitation parameter. The average NMSE of the proposed algorithm is much smaller than that of the conventional one.

Table 1. Dimension of each speech analysis/synthesis method for HMM training.

	STRAIGHT	Conv.	Prop.
LSF	24+Δ+ΔΔ	24+Δ+ΔΔ	24+Δ+ΔΔ
Excitation	5+Δ+ΔΔ	22+Δ+ΔΔ	22+Δ+ΔΔ
F0	1+Δ+ΔΔ	1+Δ+ΔΔ	1+Δ+ΔΔ
Energy	1+Δ+ΔΔ	1+Δ+ΔΔ	1+Δ+ΔΔ

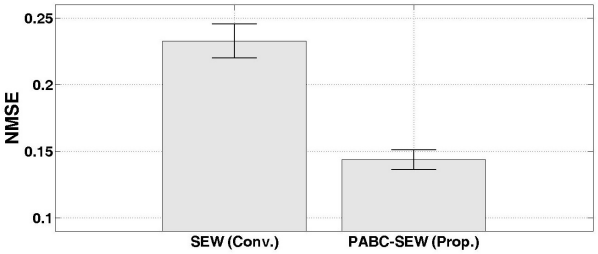

Fig. 5. NMSE of excitation parameter generated from context-dependent HMMs.

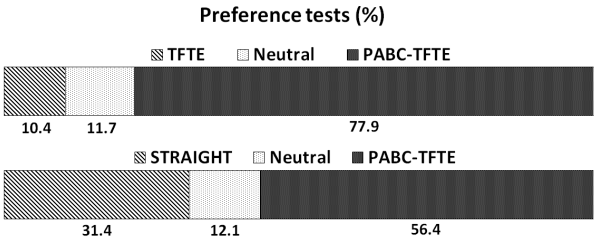

Fig. 6. Results of preference tests.

From the result, it is clear that the PABC-TFTE has an advantage over the conventional one while training the parameters with the HMMs. Furthermore, the large confidence interval of the conventional algorithm implies that the reconstructed excitation contains many frames with large errors, which results in the degradation of naturalness or inconsistent results.

4.3. Subjective Test Result

The perceptual quality of the proposed algorithm is also evaluated by an A/B preference listening test. The proposed PABC-TFTE-based system is compared to the conventional TFTE-based and STRAIGHT-based systems. Note that STRAIGHT is known as the state-of-the-art speech analysis/synthesis algorithm in the recent study of HMM-based TTS systems [8]. In the test, twelve experienced listeners were asked to make a quality judgment in an acoustically isolated room with Sennheiser HD650 headphones. The preference test results are shown in Fig. 6. The test results verify that the perceptual quality of the proposed algorithm is much better than that of the conventional one.

5. CONCLUSION

Improved TFTE modeling for a statistical parametric speech synthesis has been proposed. To overcome the drawbacks of the vocoding techniques of the conventional approaches, we adopt the PABC-based parameterization method of conventional TFTE modeling. The proposed PABC-TFTE significantly reduces the redundancy of the excitation parameters; thus, the excitation signal can be regenerated with several coefficients of the PABC-SEW and REW. As most of the remaining PABC-SEWs follow the normal distribution, they are easily modeled by a single Gaussian. As a result, training can be successfully accomplished regardless of the time-varying dimensions of the parameters.

6. REFERENCES

[1] H. Zen, K, Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Communication,* vol. 51, no. 11, pp. 1039-1064, 2009.

[2] L.H. Ling, D. Li, and Y. Dong, "Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 21, no. 10, pp. 2129-2139, 2013.

[3] H. Zen, A. Senior, and M. Schuster,"Statistical parametric speech synthesis using deep neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 7962-7966, 2013.

[4] S. Kang, X. Qian, and H. Meng,"Multi-distribution deep belief network for speech synthesis," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 8012-8016, 2013.

[5] Y. Qian, et al.,"On the training aspects of deep neural network (DNN) for parametric TTS synthesis," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 3829-3833, 2014.

[6] T. Yoshimura, et al,,"Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-Based Speech Synthesis," in *European Conference on Speech Communication and Technology*, 1999.

[7] T. Yoshimura, et al.,"Mixed-excitation for HMM-based speech synthesis," in *European Conference on Speech Communication and Technology*, 2001.

[8] H. Zen, et al., "Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," in *IEICE transactions on information and systems*, vol. 90, no.1, pp. 325-333, 2007.

[9] A. McCree, and T. Barnwell,"A mixed excitation LPC vocoder model for low bit rate speech coding," in *IEEE Transactions on Speech Audio Processing,* vol. 3, no.4, 1995.

[10] K. Kawahara, I. Masuda-Katsuse, and A. Cheveibne,"Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *1997 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 1997.

[11] J. Sung, et al.,"Excitation modeling based on waveform interpolation for HMM-based speech synthesis," in *INTERSPEECH*, 2010.

[12] C.S. Jung, Y.S. Joo, and H.G. Kang, "Waveform interpolation-based speech analysis/synthesis for HMM-based TTS systems," *IEEE Signal Processing Letters,* IEEE, vol. 19, no. 12, pp. 809-812, 2012.

[13] E. Choy, "Waveform Interpolation Speech Coder at 4kb/s," Master of Engineering, McGill Univ., Dept. Elect. Eng., Montreal, QC, Canada, 1998.

[14] J.C. Hardwick, and J.S. Lim,"A 4.8 kbps multi-band excitation speech coder," in *1988 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 374-377, 1988.

[15] H. Sakoe, and S. Chiba,"Dynamic programming algorithm optimization for spoken word recognition," in *IEEE Transactions on Acoustic, Speech and Signal Processing*, IEEE, vol. 26, no. 1, pp. 42-49, 1978.

[16] K. Tokuda, et al.,"The HMM-based speech synthesis system (HTS)," Available: http://hts.ics.nitech/ac.jp.