# Deep Neural Network-Based Statistical Parametric Speech Synthesis System Using Improved Time-Frequency Trajectory Excitation Model

*Eunwoo Song and Hong-Goo Kang*

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

sewplay@dsp.yonsei.ac.kr

## Abstract

This paper proposes a deep neural network (DNN)-based statistical parametric speech synthesis system using an improved time-frequency trajectory excitation (ITFTE) model. The ITFTE model, which efficiently reduces the parametric redundancy of a TFTE model, improved the perceptual quality of the vocoding process and the estimation accuracy of the training process. However, there remain problems related to training ITFTE parameters in a hidden Markov model (HMM) framework, such as inefficiency of representing cross-dimensional correlations between ITFTE parameters, over-smoothed outputs caused by statistical averaging, and an over-fitted model due to a decision tree-based state clustering paradigm. To alleviate these limitations, a centralized DNN replaces the decision trees of the HMM training process. Analysis of trainability confirms that the DNN training process improves the model accuracy, which results in improved perceptual quality of synthesized speech. Objective and subjective test results also verify that the proposed system performs better than the conventional HMM-based system.

**Index Terms**: Statistical parametric speech synthesis system, deep neural network (DNN), time-frequency trajectory excitation (TFTE)

## 1. Introduction

In statistical parametric speech synthesis systems, an efficient vocoding technique is needed to generate the natural qualities of synthesized speech. To reduce the buzziness of the conventional pulse-or-noise (PoN) model [1], various approaches have been adopted to accurately model the excitation signal. By dividing the whole frequency band into several fixed sub-bands [2, 3], the excitation signal of each sub-band was represented by either PoN or band aperiodicities (BAPs) [4, 5]. The quality of synthesized speech was significantly improved, as mixed excitation is a more generalized representation than the PoN model. However, its perceptual quality was still buzzy and unnatural because fixing the boundary of each frequency band could not fully represent the time-varying periodicity of various types of phonetic information.

To alleviate the aforementioned limitations, a time-frequency trajectory excitation (TFTE)-based speech synthesis system was proposed [6, 7]. By decomposing pitch-dependent excitation into a slowly evolving waveform (SEW) and rapidly evolving waveform (REW) [8], the TFTE extracts the voicing ratio of an excitation signal in every frequency bin.

Although the TFTE-based vocoding technique provided high-quality synthesized speech, the full merits of SEW/REW decomposition could not be utilized in the model training process because its parametric dimension varies depending on the length of the pitch interval. In our previous work [9], we proposed an improved TFTE (ITFTE) model for the hidden Markov model (HMM)-based speech synthesis system, which overcame the conventional problem of time-varying feature dimensions. By using the characteristic of time-frequency correlation in SEW magnitude, the whole frequency band of SEW could be represented only by several fixed number of discrete cosine transform (DCT) coefficients. These coefficients were used in the HMM training process; on the other hand, the remaining coefficients were stochastically generated by Gaussian random variables. Consequently, the training process avoided problems associated with dimensional variation.

In this paper, to further improve the whole framework, we propose a deep neural network (DNN)-based statistical training process using the ITFTE model. Since SEW and REW represent a voicing ratio of phonetic information on the time-frequency surface, it is very important to accurately model the dynamic nature of evolving characteristics during the training process. However, there are several limitations on training ITFTE parameters with HMMs. Firstly, HMMs cannot fully model cross-dimensional correlations between ITFTE parameters because each state is modeled by a single Gaussian, diagonal covariance output distribution. Secondly, a statistical averaging process embedded in the HMM training process creates an over-smoothing problem, which results in the loss of time-varying characteristics of ITFTE parameters. Furthermore, having prohibitively large tree and separating training data can lead to an over-fitting problem [10]. As a result, the detailed dynamic characteristics of ITFTE parameters are lost during the training process, thereby degrading the quality of synthesized speech.

To address these limitations, a deep neural network (DNN) architecture is used to build a non-linear function that maps contextual information to corresponding ITFTE parameters. It is well known that the DNN structure has advantages such as: the ability to represent complicated functions of features compactly; a centralized network that can model all training data without data fragmentation; a deep-layered and hierarchical structure that can identify highly non-linear relationships between input and output features [10–12]. By replacing conventional decision tree-clustered context-dependent HMMs with a DNN architecture, the proposed system improves the accuracy of training ITFTE parameters. The trainability of the proposed system is compared in detail with the conventional HMM-based approach. The findings confirm that the proposed system not only reduces estimation errors during training, but also provides more natural quality to the synthesized speech. Experimental results also verify that the proposed system achieves superior objective and subjective speech quality to that of the conventional method.

# 2. Speech Synthesis Using Time-Frequency Trajectory Excitation

## 2.1. Time-frequency trajectory excitation

TFTE exploits a time-frequency surface to represent the voicing characteristics of an excitation signal. Let $u(n, \phi)$ denote a periodic function with $\phi$ 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 $\phi$ is defined as $\phi(m) = 2\pi m/P(n)$ with a pitch period $P(n)$, and $A_k(n)$ and $B_k(n)$ denote the $k$-th discrete-time Fourier series coefficients of the excitation signal [8].

To extract the voicing ratio of each individual 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:

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

The SEW and REW therefore represent the periodic and remaining noisy components of TFTE in each frequency bin, respectively.

## 2.2. ITFTE modeling for statistical parametric speech synthesis system

The TFTE parameters should be appropriately adjusted before being used for training. Note that the number of parameters to be modeled in each pitch epoch varies because of the pitch-dependent analysis paradigm.

The periodic component, SEW, can be efficiently modeled by a fixed-number of DCT coefficients with stochastically generated random variables. The SEW magnitude is first divided into $K$ number of frequency sub-blocks,

$$\begin{bmatrix} c_{k,1} \\ \vdots \\ c_{k,J_k} \end{bmatrix}^T = \begin{bmatrix} u_{SEW}(n, J_{k-1}+1) \\ \vdots \\ u_{SEW}(n, J_{k-1}+J_k) \end{bmatrix}^T, \ 1 \le k \le K, \quad (4)$$

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

$$\sum_{k=1}^{K} J_k = P(n)/2, \quad (5)$$

where $P(n)/2$ is the length of the SEW. When dividing the whole frequency band into sub-blocks, logarithm-scale (*e.g.* mel-scale) division is recommended. Note that high resolution of low-frequency regions improves the perceived quality of synthesized speech, compared to that of equally divided sub-blocks. Each SEW sub-block is then transformed with the DCT:

$$C_{k,m} = \frac{1}{J_k}\sum_{j=1}^{J_k} c_{k,j}\cos\left(\frac{\pi}{J_k}(j-0.5)(m-1)\right), \quad (6)$$
$$1 \le m \le J_k$$

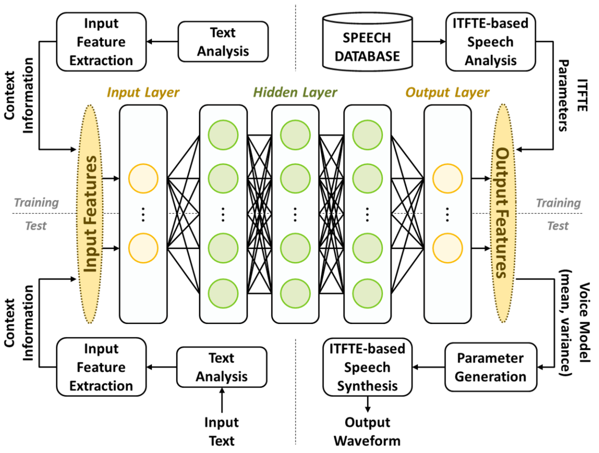


Figure 1: A framework of DNN-based speech synthesis system using the ITFTE model.

where $C_{k,m}$ represents the $m$-th DCT coefficient of the $k$-th sub-block. As the DCT has good decorrelation and energy compactness properties [13], most sub-block information of the SEW magnitude is concentrated within the first few coefficients. The remaining coefficients can be modeled by a Gaussian mixture model (GMM), since they follow a normal distribution [9]. As a result, the entire frequency band information of SEW magnitude can be trained with only several fixed number of DCT sub-block coefficients, while the remaining coefficients are generated by Gaussian random variables in the synthesis step.

The noisy component, REW, is modeled by a power contour estimation method because its perceptual quality is not much different [8]. Typically, a fixed number of Legendre orthonormal polynomial coefficients are used to parameterize the REW components.

# 3. DNN-Based Speech Synthesis System Using ITFTE Model

This section describes a DNN-based speech synthesis system using the ITFTE modeling technique. In the conventional HMM-based speech synthesis system, acoustic parameters are clustered by decision trees and trained by corresponding HMMs. In contrast, the proposed system replaces the thousands of GMM leaf nodes of decision trees with a centralized DNN that is subsequently trained to construct a mapping function from the contextual information to the corresponding acoustic parameters. To verify the advantages of the proposed DNN-based system, we compare the trainability of the ITFTE parameters to that of the conventional HMM-based system.

## 3.1. Speech synthesis system based on a deep architecture

Figure 1 illustrates a framework of the DNN-based speech synthesis system. In the training step, after analyzing the contextual information of input texts, they are converted to a sequence of input features. The input features include binary features (*e.g.* the previous, current, next phoneme identity, and the position of the current phoneme in the current syllable) and numerical features (*e.g.* the number of phonemes/syllables/words in the current syllable/word/utterance, the position of the current syllable/word in the current word/utterance, and the durations of current phoneme). The output features contain acoustic parameters (*e.g.* line spectral frequencies, fundamental frequency, energy, and parameterized SEW and REW coefficients) with their time dynamics [14]. The weights of DNN are then trained to minimize the mean square error between target and estimated outputs with regard to given inputs.

In the test step, the contextual information is first converted to input features, and then the output features are estimated by the trained DNN and the input features. By setting the estimated output feature vector as a mean vector and pre-computed global variances of output feature vectors, continuous trajectories of the speech parameters can be generated by means of a speech parameter generation algorithm [15]. Finally, the ITFTE synthesis module synthesizes a speech waveform with the generated parameters.

### 3.2. Analysis of the trainability of DNN-based speech synthesis system using ITFTE model

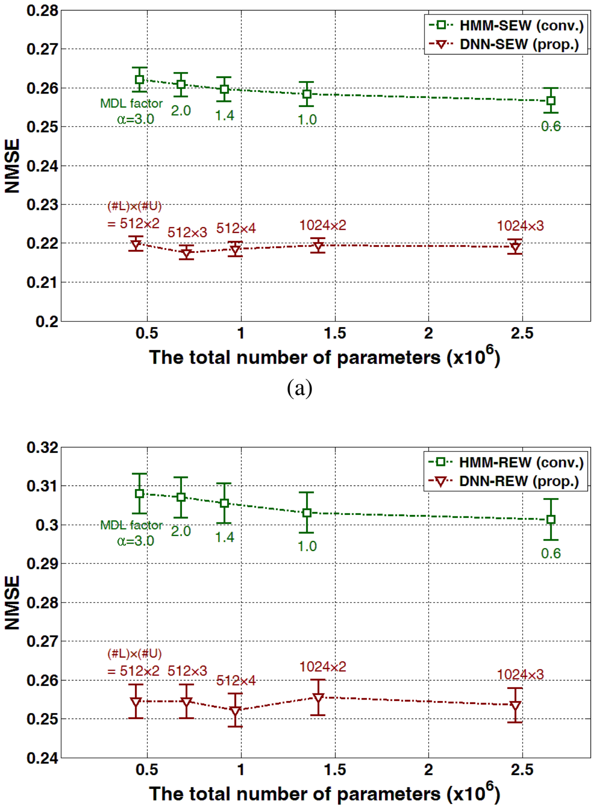
This section describes the advantages of utilizing the ITFTE model for the DNN-based speech synthesis system. To evaluate the effectiveness of proposed system compared to the conventional HMM-based method, the normalized root mean square error (NMSE) is measured. The NMSE is defined as the normalized error between the original and generated excitation parameter:

$$NMSE = \frac{1}{N}\sum_{n=1}^{N}\sqrt{\frac{\sum_{k=1}^{K}\left(x_{ori}(n,k)-x_{gen}(n,k)\right)^2}{\sum_{k=1}^{K}\left(x_{ori}(n,k)\right)^2}}, \quad (7)$$

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 excitation parameters extracted from the original speech and those generated by the trained HMM/DNN, respectively.

Figure 2 represents the mean NMSE with a 95% confidence interval for each system, depending on the total number of parameters. In the DNN-based system, the model size is controlled by adjusting the number of layers (#L) and the number of units (#U). In the HMM-based method, on the other hand, model size is controlled by changing a scale factor of the minimum description length (MDL) criteria [16]. The duration in both system is modeled by HMMs and the dynamic time warping technique is used to compensate for the durational mismatch between the original and generated signals [17]. More setup details are shown in section 4.1.

The results represent the effectiveness of each system in several ways. Firstly, the average NMSE of the HMM-based system is large compared to that of the DNN-based one. Furthermore, the parameterized REW contains much larger estimation error than the parameterized SEW, since the dynamically varying characteristics of REW are over-smoothed by statistical averaging as part of the HMM training process. Secondly, the wider confidence interval of the HMM-based system implies that the excitation signal contains many frames with large errors, which would be expected to degrade naturalness or produce inconsistent results in the synthesis step. On the contrary, the proposed DNN-based system reduces estimation error in the training process. It is therefore expected to provide more natural quality of synthesized speech than the conventional HMM-based system, of which results are confirmed in the next section.



(a)

(b)

Figure 2: NMSE of excitation parameters generated from HMM-based (conv.) and DNN-based (prop.) system: (a) SEW and (b) REW coefficients

## 4. Experiments

### 4.1. Experimental setup

Phonetically and prosodically balanced speech data were used for training the HMM-based and DNN-based speech synthesis systems. In total, 2,700 utterances (around 3.5 hours) were used for training, and 100 utterances were used for validation (only for the DNN-based system). Extra 100 utterances not included in the training and validation were used for evaluation. The corpus was recorded by a male professional speaker. The speech signals were sampled at 16 kHz, and each sample was quantized by 16 bits. A grapheme-to-phoneme (G2P) converter was developed by following rules of the Korean standard pronunciation grammar and the context information-labeling program.

In the analysis step, the frame length is set to 20 ms, and the spectral and excitation parameters are extracted every 5 ms. The 24-dimensional line spectral frequencies (LSFs) are extracted for the spectral parameter, whereas 18-dimensional SEW DCT sub-block coefficients and 4-dimensional REW polynomial coefficients are extracted for the excitation parameters. Logarithmic fundamental frequency (log-F0) and energy are also extracted for training the HMM/DNN. Table 1 summarizes the dimensions of each parameter. In the synthesis step, all parameters are generated by a speech parameter generation algorithm. The generated SEW DCT sub-block coefficients reconstruct the SEW magnitude with Gaussian random variables, whereas the REW magnitude is reconstructed by the generated polynomial coefficients. The TFTE is then reconstructed from the SEW and

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

| | **STRAIGHT** | **ITFTE** |
|---|---|---|
| LSF | 24+Δ+ΔΔ | 24+Δ+ΔΔ |
| Excitation | 22+Δ+ΔΔ | 22+Δ+ΔΔ |
| Log-F0 | 1+Δ+ΔΔ | 1+Δ+ΔΔ |
| Energy | 1+Δ+ΔΔ | 1+Δ+ΔΔ |
| V/UV | 1 | None |

REW with its pitch period. The SEW phase is extracted from a recorded speech, whereas the REW phase is randomly selected. Finally, the single pitch-based speech signal is synthesized by the generated LSFs and TFTE.

In the DNN-based speech synthesis system, the input feature vector includes 210-dimensional contextual information, which consists of 203 binary features for categorical linguistic contexts and 7 numerical features for numerical linguistic contexts. The output feature vector contains 144-dimensional acoustic parameters that consist of LSFs, parameterized SEW and REW coefficients, log-F0, and energy. The time dynamics of these parameters are also included for the speech parameter generation algorithm. Before training, both input and output features are normalized: zero-mean unit-variance normalization is used for input features; on the other hand, minimum-maximum (from $0.01$ to $0.99$) normalization is used for output features. In the training, the weights are initialized randomly and trained by using a back-propagation procedure based on a mini-batch stochastic gradient descent algorithm [18].

To evaluate the performance of proposed system, objective and subjective test results are compared to those of the conventional system based on HMM-ITFTE. In the subjective test, additional DNN-based system using STRAIGHT is also included. Note that STRAIGHT is well known as the high quality speech analysis/synthesis algorithm in the recent study of speech synthesis systems [5]. In the HMM-ITFTE system, a context-dependent speech synthesis system is constructed (setup details are given in [7]). In order to ensure a fair comparison with the DNN-based system, the model size is controlled by adjusting a scale factor of the MDL criteria. In the DNN-STRAIGHT system, the experimental setup is same as the proposed DNN-based system, except for the output features. The acoustic parameters are changed to be suitable for STRAIGHT, including 24-dimensional LSFs, 22-dimensional BAPs, log-F0, energy, and their time dynamics. The binary value of voiced/unvoiced (V/UV) information is added to the output features, and log-F0 values are interpolated for the unvoiced frames [10].

### 4.2. Objective test results

To evaluate the objective quality of synthesized speech, we compare distortions in acoustic parameters obtained from original speech with those estimated by DNN/HMM systems. The metrics for measuring distortion are log-spectral distance (LSD) for LSFs in dB, RMSE for F0 in Hz, and NMSE for SEW and REW. The test results for different architecture setups of DNN systems and MDL factors of HMM systems are shown in Tables 2 and 3. From the results, it is clear that the ITFTE parameters generated by the DNN-based system contain smaller estimation errors than those generated by the HMM-based method. The F0 RMSE of the DNN-based system is also smaller than that of the HMM-based one, which results in more accurate reconstruction of whole frequency SEW and REW magnitudes. Although the results on LSD are similar in both systems, it is clear that the DNN-based system has advantages over the HMM-based one in terms of modeling the ITFTE-related parameters. Moreover, since the accuracy of these parameters is highly related to the quality of synthesized speech, the perceptual quality of synthesized speech from the DNN-based system is expected to be better than that from the HMM-based one.

Table 2: Test results for different MDL factors (with model size) of the conventional HMM-ITFTE system.

| HMM-<br>ITFTE | LSD<br>(dB) | F0 RMSE<br>(Hz) | SEW<br>NMSE | REW<br>NMSE |
|---|---|---|---|---|
| $\alpha$=3.0 (0.44M) | 3.227 | 16.788 | 0.262 | 0.308 |
| $\alpha$=2.0 (0.68M) | 3.188 | 16.732 | 0.261 | 0.307 |
| $\alpha$=1.0 (1.35M) | 3.129 | 16.401 | 0.258 | 0.303 |
| $\alpha$=0.6 (2.65M) | 3.098 | 16.759 | 0.257 | 0.301 |

Table 3: Test results for different architectures (with model size) of the proposed DNN-ITFTE system.

| DNN-<br>ITFTE | LSD<br>(dB) | F0 RMSE<br>(Hz) | SEW<br>NMSE | REW<br>NMSE |
|---|---|---|---|---|
| 512×2 (0.46M) | 3.240 | 14.748 | 0.220 | 0.255 |
| 512×3 (0.71M) | 3.192 | 13.218 | 0.218 | 0.254 |
| 1024×2 (1.41M) | 3.207 | 14.477 | 0.219 | 0.256 |
| 1024×3 (2.46M) | 3.189 | 15.766 | 0.219 | 0.254 |

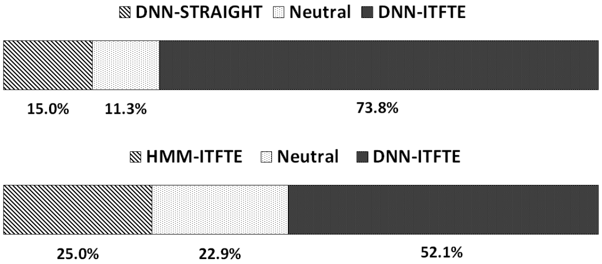


Figure 3: Results of preference tests (%).

### 4.3. Subjective test results

The perceptual quality of the proposed system is evaluated by performing an A/B preference listening test. The proposed system is compared with the conventional systems using HMM-ITFTE and DNN-STRAIGHT. In the test, twelve experienced listeners are asked to provide quality judgments. Twenty utterances are randomly selected from the evaluation set, then they are synthesized by the HMM-based system ($\alpha = 0.6$) and the DNN-based systems ($1024 \times 3$). Figure 3 depicts the results of the preference test, which verifies that the proposed system provides much higher perceptual quality than that of conventional systems.

## 5. Conclusions

A DNN-based statistical parametric speech synthesis system using the ITFTE model has been proposed. To overcome the limitations of conventional HMM-based speech synthesis systems, a centralized DNN model was introduced to replace the decision trees of the HMM training process. The proposed DNN training process improved model accuracy; thus, significantly reducing the estimation error while generating ITFTE parameters in the synthesis step. Subjective listening tests also confirmed the superiority of the proposed system over the conventional ones.

## 6. 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*, 1999.

[2] A. V. McCree and T. P. Barnwell III, “A mixed excitation LPC vocoder model for low bit rate speech coding,” *IEEE Trans. Speech and Audio Process.*, vol. 3, no. 4, pp. 242–250, 1995.

[3] H. Kawahara, “Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited,” in *Proc. ICASSP*, 1997.

[4] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Mixed excitation for HMM-based speech synthesis,” in *Proc. EUROSPEECH*, 2001.

[5] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, “Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005,” *IEICE trans. Inf. Syst.*, vol. 90, no. 1, pp. 325–333, 2007.

[6] J. S. Sung, D. H. Hong, K. H. Oh, and N. S. Kim, “Excitation modeling based on waveform interpolation for HMM-based speech synthesis.” in *Proc. INTERSPEECH*, 2010.

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

[8] E. L. Choy, “Waveform interpolation speech coder at 4 kb/s,” Ph.D. dissertation, McGill University Montreal, Canada, 1998.

[9] E. Song, Y. S. Joo, and H. G. Kang, “Improved time-frequency trajectory excitation modeling for a statistical parametric speech synthesis system,” in *Proc. ICASSP*, 2015.

[10] H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” in *Proc. ICASSP*, 2013.

[11] Y. Bengio, “Learning deep architectures for AI,” *Foundations and trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[12] Y. Qian, Y. Fan, W. Hu, and F. K. Soong, “On the training aspects of deep neural network (DNN) for parametric TTS synthesis,” in *Proc. ICASSP*, 2014.

[13] J. C. Hardwick and J. S. Lim, “A 4.8 kbps multi-band excitation speech coder,” in *Proc. ICASSP*, 1988.

[14] S. Furui, “Speaker-independent isolated word recognition using dynamic features of speech spectrum,” *IEEE Trans. Acoust., Speech Signal Process.*, vol. 34, no. 1, pp. 52–59, 1986.

[15] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” in *Proc. ICASSP*, 2000.

[16] K. Shinoda and T. Watanabe, “MDL-based context-dependent subword modeling for speech recognition,” *J. Acoust. Soc. Jpn(E)*, vol. 21, no. 2, pp. 79–86, 2000.

[17] H. Sakoe and S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition,” *IEEE Trans. Acoust., Speech Signal Process.*, vol. 26, no. 1, pp. 43–49, 1978.

[18] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning internal representations by error propagation,” DTIC Document, Tech. Rep., 1985.