Non-filter waveform generation from cepstrum using spectral phase reconstruction

Yasuhiro Hamada[1], *Nobutaka Ono*[2], *Shigeki Sagayama*[1]

[1]Meiji University, Nakano, Tokyo, Japan
[2]National Institute of Informatics / The Graduate University for Advanced Studies, Tokyo, Japan
hamada@meiji.ac.jp, onono@nii.ac.jp, sagayama@meiji.ac.jp

Abstract

This paper discusses non-filter waveform generation from cepstral features using spectral phase reconstruction as an alternative method to replace the conventional source-filter model in text-to-speech (TTS) systems. As the primary purpose of the use of filters is considered as producing a waveform from the desired spectrum shape, one possible alternative of the source-filter framework is to directly convert the designed spectrum into a waveform by utilizing a recently developed " phase reconstruction "from the power spectrogram. Given cepstral features and fundamental frequency (F_0) as desired spectrum from a TTS system, the spectrum to be heard by the listener is calculated by converting the cepstral features into a linear-scale power spectrum and multiplying with the pitch structure of F_0. The signal waveform is generated from the power spectrogram by spectral phase reconstruction. An advantageous property of the proposed method is that it is free from undesired amplitude and long time decay often caused by sharp resonances in recursive filters. In preliminary experiments, we compared temporal and gain characteristics of the synthesized speech using the proposed method and mel-log spectrum approximation (MLSA) filter. Results show the proposed method performed better than the MLSA filter in the both characteristics of the synthesized speech, and imply a desirable properties of the proposed method for speech synthesis.

Index Terms: HMM-based speech synthesis, cepstral features, non-filter, spectral phase reconstruction

1. Introduction

In the long history of speech synthesis, source-filter model has played an essential role as a simulator of the human process of speech production consisting of excitation and resonances. Source-filter model was already employed in the very early stage of the history of speech synthesis using acoustic and mechanical components such as Wolfgang von Kempelen 's work [1] in the mid-18th century. Electrical speech synthesis first introduced by Stewart in 1922 [2] and followed by a number of electrical speech synthesizers including vocal tract analog, terminal analog and formant-based synthesizers also utilized the source-filter model [3].

After they were replaced by digital calculation by computers, artificial speech waveform has been synthesized by the digital source-filter model (e.g. linear predictive coding (LPC) vocoder [4, 5], line spectrum pair (LSP)[6], etc.). Even after statistical speech synthesis was proposed in 1989 using context oriented clustering (COC) [7, 8, 9], deployed as a commercial product (" Shaberinbo " from NTT Data), and, from the 90's, sophisticated with hidden Markov model (HMM) toward better natural-sounding synthetic speech[10], the source-filter model has still been one of main components using the mel-log spectrum approximation (MLSA) filter [11] to generate waveform from cepstral features.

Apart from waveform unit concatenation methods such as PSOLA [12], source-filter model seems to be almost exclusively used for producing synthetic speech waveform from the intended and designed spectrum in the text-to-speech (TTS) framework. Our view is twofold over the reason of conventional use of source-filter model in the TTS framework. One reason is that it simulates the human speech production mechanism. The other reason is that filter has been considered as almost the only means to generate waveform from the given spectrogram supplied from the TTS part. As the ultimate purpose of the speech waveform synthesizer is to deliver the spectrogram designed in the TTS system to the listener 's brain, the waveform is essential as the means for transmitting the " target " spectrum to the listener. For this purpose, the source-filter model has long been considered to be the best solution. An alternative solution for this problem can be brought from " phase reconstruction " [13, 14] without using filters.

We propose a spectral phase reconstruction, instead of using filter, to generate waveform from power spectrum. An advantageous property of the proposed method is that it is free from undesired amplitude and long time decay often caused by sharp resonances in recursive filters.

2. Non-filter speech synthesis

2.1. Cepstral features and F_0 generation based on HMM

We partly use a typical HMM-based speech synthesis system [15, 16] statistically trained for generating the parameter vector sequence of mel-frequency cepstrum coefficients (MFCCs) and F_0. This vector sequence is assumed to represent the target spectral time sequence (spectrogram) designed to be heard by the listener.

2.2. Conversion from cepstrum to spectrum

The given MFCCs are converted to linear-scaled spectrum as follows [17].

$$H(\omega) = s_\gamma^{-1}\left(\sum_{m=0}^{M} \tilde{c_\gamma}(m) e^{-j\tilde{\omega}t}\right) \tag{1}$$

where

$$s_\gamma^{-1}(\omega) = \begin{cases} (1+\gamma\omega)^{1/\gamma}, & 0 < |\gamma| \leq 1 \\ \exp\omega, & \gamma = 0 \end{cases} \tag{2}$$

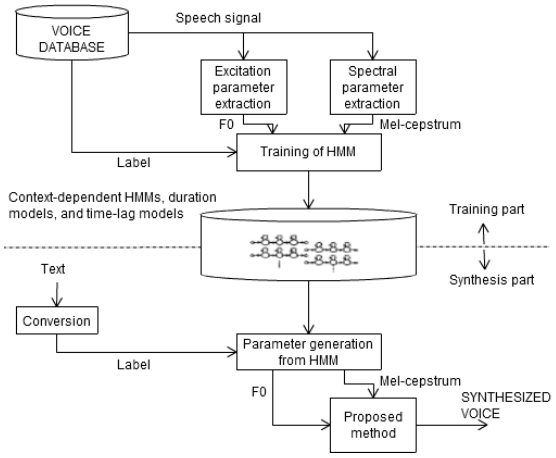

Figure 1: *Overview of proposed HMM-based synthesis system*

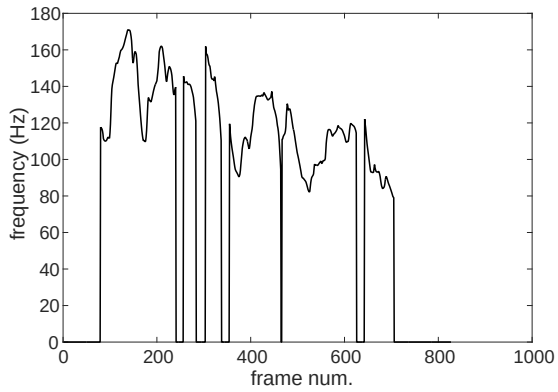

Figure 2: *A sample of F_0 trajectory of speech*

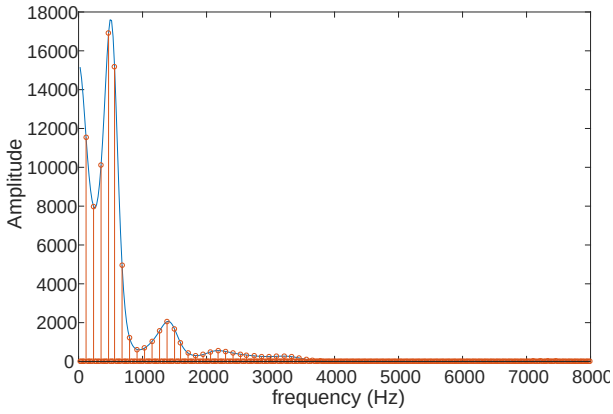

Figure 3: *Harmonic components of a spectrum*

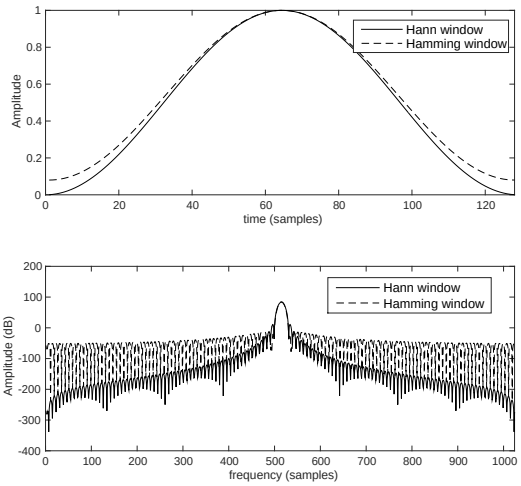

Figure 4: *Hann window (top) and spectrum of Hann window (bottom)*

$$\tilde{\omega} = \tan^{-1} \frac{(1-\alpha^2)\sin\omega}{(1+\alpha^2)\cos\omega - 2\alpha} \tag{3}$$

$H(\omega)$ denotes the spectrum, s_γ^{-1} is the inverse generalized log function, $\tilde{c}_\gamma(m)$ is the mel-generalized cepstral coefficients and α denotes the frequency compression parameter. The generalized mel-cepstrum is equivalent to the cepstrum with $\gamma = 0$ and AR coefficient with $\gamma = -1$.

2.3. Adding F_0 components

Using F_0 generated by the HMM, harmonic components can be added to spectrum $H(\omega)$ by sampling $H(\omega)$ at integral multiples of F_0. Fig. 2 shows a F_0 trajectory of speech. Fig. 3 shows a sample spectrum with F_0 being $120Hz$ at the frame number 100.

In the case of voiced speech sound, F_0-harmonic components are convolved with the spectrum domain representation of the Hann window, $W(\omega)$, as follows to obtain the target spectrum. That can be described as harmonic components, because the spectrum sidelobe of Hann window decreases rapidly as shown in fig. 4

$$w(n) = 0.5(1 - \cos(\frac{2\pi n}{N-1})), n = 1, 2, ...N \tag{4}$$

$$W(\omega) = \exp(i\omega\frac{N-1}{2})\frac{\sin(N\omega/2)}{\sin(\omega/2)} \tag{5}$$

$$X(\omega) = H(\omega) * W(\omega) \tag{6}$$

Fig. 4 shows the Hann window (top) and the spectrum of the Hann window (bottom).

Fig. 5 shows a sample spectrogram of $H(\omega)$ convoluted with $W(\omega)$

2.4. Spectral phase reconstruction

Speech waveform is synthesized from the power spectrum using spectral phase reconstruction [13, 14] where the spectral phase is updated iteratively by short-time Fourier transform (STFT) and inverse STFT (Fig. 6).

The process of spectral phase reconstruction is described as follows.

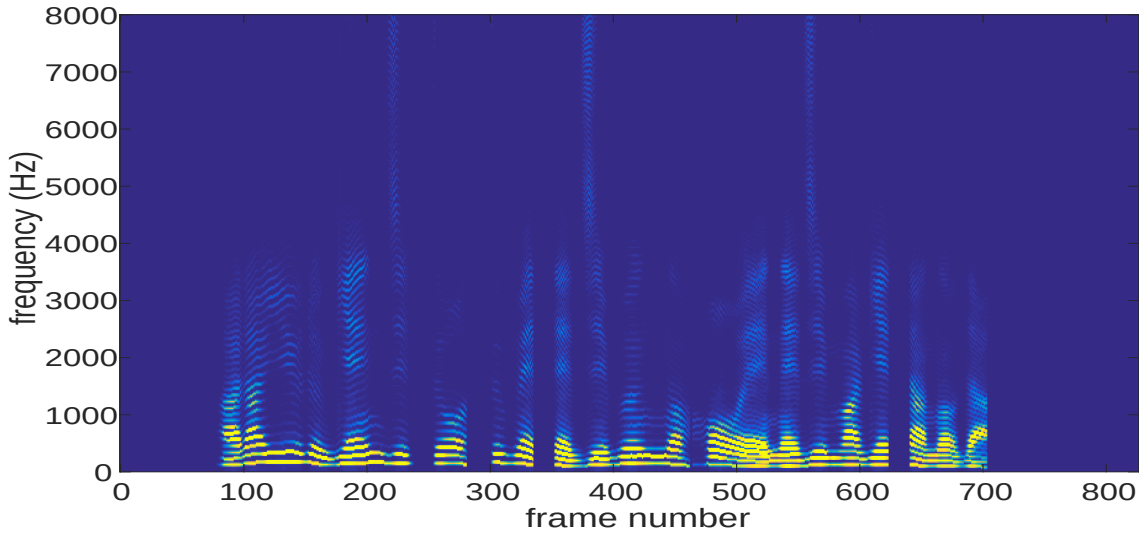

Figure 5: *Synthetic spectrogram of convolved $H(\omega)$ and $W(\omega)$*

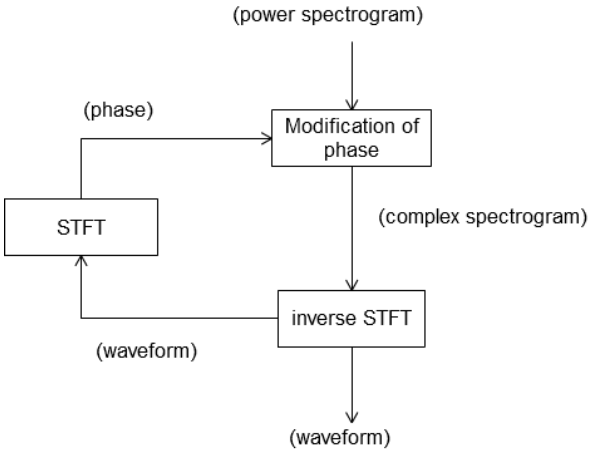

Figure 6: *Algorithm of spectral phase reconstruction*

1. Complex spectrogram $X^{(0)}[mS,k]$ is calculated. Initial phases are set to the given power spectrogram $|X[mS,k]|^2$ (where S equal frame shift).
2. Waveform $x^{(i)}[n]$ is generated by calculating as follows.

$$x^{(i)}[n] = \begin{array}{c} \frac{1}{N}\sum_{m=-\infty}^{\infty} w[n-mS] \\ \sum_{k=0}^{N-1} X^{(i)}[mS,k]e^{-j\frac{2\pi}{N}(n-mS)k} \end{array} \quad (7)$$

   where w is the window function.
3. Complex spectrogram $\hat{X}^{(i)}[mS,k]$ is calculated from the STFT of $x^{(i)}[n]$.

$$\hat{X}^{(i)}[mS,k] = \sum_{n=-\infty}^{\infty} x^{(i)}[n]w[n-mS]e^{-j\frac{2\pi}{N}(n-mS)k} \quad (8)$$

4. The given power spectrogram $|X[mS,k]|^2$ are combined with the phase of the complex spectrogram $\hat{X}^{(i)}[mS,k]$.

$$X^{(i+1)}[mS,k] = \hat{X^{(i)}}[mS,k]\frac{|X[mS,k]|}{|\hat{X}^{(i)}[mS,k]|} \quad (9)$$

5. Iterate step 2 – step 4.

3. Evaluation of temporal and gain characteristics

3.1. Problems in temporal and gain characteristics of source-filter model

Generally in vocoders, naturalness of re-synthesized speech is degraded when F_0 is modified from the original F_0. Major factors of such quality degradation are considered to be related to temporal and gain characteristics of the synthesized speech. When using a recursive filter such as the MLSA filter, if some harmonic components of F_0 overlap a formant of a high Q value, it often causes an undesired sharp resonance resulting in a large amplitude and lengthened decay spanning a few frames. Furthermore, because gain of an output signal is linearly proportional to the Q value, gain fluctuates according to F_0. As the proposed method is not a recursive filter and free from sharp resonances, these problems can be reduced. For these reasons, we experimentally compared the temporal and gain characteristics of the proposed and MLSA methods.

3.2. Experimental conditions

As the speech materials, we synthesized speech using cepstrum characteristics and F_0 generated by HTS [16] from five texts. The cepstral parameters were set as: $\gamma = 1.0$ and $\alpha = 0.55$. The STFT was performed with a sampling frequency of 48000 Hz, frame shift of 5 ms and frame length of 40 ms. In the spectral phase reconstruction, the phase estimation step was iterated fifty times for the synthetic speech duration ranging from 3.35 to 5.25 s. F_0 was modified from 0.8 to 1.2 times of the original with the interval of 0.05.

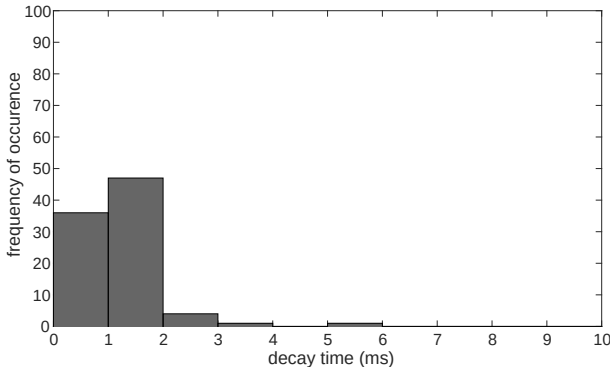

Figure 7: *Temporal characteristics of the MLSA filter*

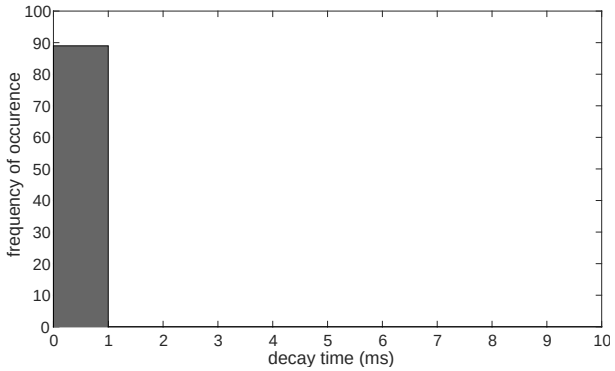

Figure 8: *Temporal characteristics of the proposed method*

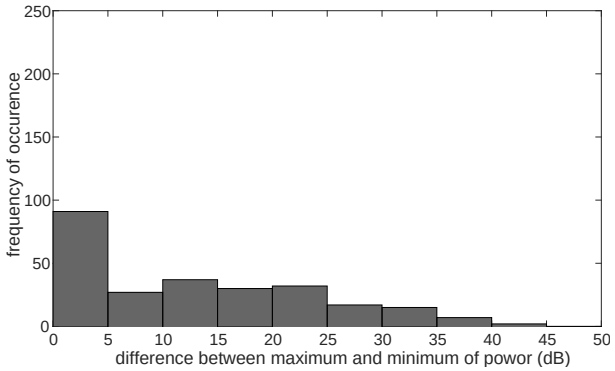

Figure 9: *Gain characteristics of the MLSA filter*

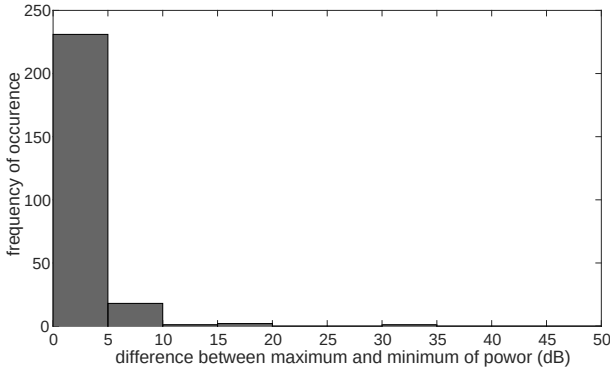

Figure 10: *Gain characteristics of the proposed method*

3.3. Evaluation of temporal characteristics and results

In the filter case, 30 ms of excitation signal was fed to the filter followed by null signal to synthesize the speech sound, while, in the phase reconstruction case, 30 ms of synthesized spectrum followed by null spectrum was converted to synthetic speech waveform. The decay times were measured in each frame for different F_0s for each speech. The decay time was defined as the time until the power became 30 dB lower, where power was the sum of squares of amplitude. Figs. 7 and 8 show time characteristics in the histogram style.

The results show that the proposed method improved the time characteristics of the synthetic speech by the MLSA method.

3.4. Evaluation of gain characteristics and results

Similarly as in the time characteristics investigation, F_0 was modified to investigate the output power of each frame in the voiced interval. Figs. 9 and 10 show gain characteristics in histogram.

The results show that the proposed method improved the gain characteristics of the synthetic speech by the MLSA method.

3.5. Naturalness

To investigate the naturalness of synthesized speech, listening tests were conducted. The stimuli consisted 90 synthetic speech with modified F_0 described in section 3.2 and presented to the subjects through an audio interface (QUAD-CAPTURE, ROLAND) and headphones (MDR-CD900ST, SONY). Seven normal-hearing Japanese (average age 21 years ranging 20–22) participated in the listening tests.

Fig. 11 shows the result of the naturalness test. The mean opinion score of the speech synthesized by MLSA filter was 2.59 while that of phase reconstruction attained 3.02.

3.6. Discussion

As the proposed method is not based on recursive filter, it has no factor causing degradation in time and gain characteristics and actually showed favorable characteristics compared with the MLSA filter. From the result of the preliminary listening test, the proposed method showed higher naturalness of the synthetic speech than that of the MLSA method. The authors 'subjective impression at the synthesized speech using the MLSA filter was " buzzy " possibly because of filter, while the proposed method sounded slightly smoother and clearer.

4. Conclusion

In this study, a novel approach of speech waveform synthesis was proposed by applying spectral phase reconstruction to speech synthesis from cepstrum characteristics and F_0 instead of the conventional source-filter model. Cepstral features generated from HMM were converted to linear-scale spectrum, sampled at the integral multiples of F_0 and convolved with the spectrum of Hann window to design the target spectrum to be heard by the listener. Then, spectral phase reconstruction technique was applied to the target spectrum to generate the waveform. In the experimental evaluation of temporal and gain character-

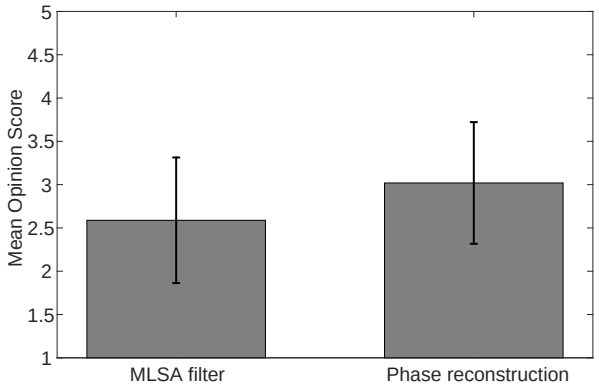

Figure 11: *Naturalness of the synthesis speech*

istics of the proposed and MLSA methods, the results showed that the proposed method gave temporal and gain characteristics better than the MLSA method, and implied that the proposed method is an effective method for speech synthesis. Future research include more detailed evaluation of synthetic speech by the non-filter approach, TTS system implementation in combination with the conventional HMM-based cepstrum and F_0 synthesizer, and a new statistical TTS framework in a new parametric domain.

5. Acknowledgments

This research was partly supported by the Grant-in-Aid for Scientific Research (A)(No. 26240025).

6. References

[1] W. von Kempelen, *Mechanismus der menschlichen Sprache nebst der Beschreibung seiner sprechenden Maschine*, 1791.

[2] John Q. Stewart, "An electrical analogue of the vocal organs," *Nature*, vol. 110, no. 7, pp. 311–312, 1922.

[3] D. H. Klatt, "Review of text-to-speech conversion for English." *The Journal of the Acoustical Society of America*, vol. 82, no. 3, pp. 737–793, 1987.

[4] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," *Reports of the 6th Int.Cong.Acoust*, no. C-5-5, 1968.

[5] B. Atal and M. Schroeder, "Predictive coding of speech signals," *Reports of the 6th Int.Cong.Acoust*, no. C-5-4, 1968.

[6] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signals," *Acoustical Society of America*, no. 57, 1975.

[7] S. Nakajima, H. Hamada, and S. Sagayama, "Speech Synthesis Method Based on Automatic Unit Generation," *Proc. ASJ Spring Meeting*, pp. 211–212, 1987.

[8] S. Nakajima and H. Hamada, "Speech synthesis method based on context oriented clustering," *IEICE Transactions*, vol. 72, no. 11, pp. 1174–1179, 1989.

[9] N. I. T. Corporation, "A high-quality text-to-speech synthesizer bord 「SHABERINBO HG」," *Journal of Acoustical Society of Japan*, vol. 49, no. 12, p. 881, 1993.

[10] T. Mashiko, K. Tokuda, T. Kobayashi, and S. Imai, "HMM-Based Speech Synthesis Using Dynamic Features," *IEEE Technical Report*, vol. 79, no. 12, pp. 2184–2190, 1996.

[11] S. Imai, K. Sumita, and F. Chieko, "Mel log spectrum approximation (mlsa) filter for speech synthesis," *IEICE Technical Report*, vol. 2, no. 66, pp. 122–129, 1989.

[12] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5-6, pp. 453–467, 1990.

[13] D. Griffin, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[14] J. Le Roux, N. Ono, and S. Sagayama, "Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction," *Proc. SAPA*, no. Sapa, pp. 23–28, 2008.

[15] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[16] H. Zen, T. Nose, J. Yamagishi, and S. Sako, "The HMM-based speech synthesis system (HTS) version 2.0." *SSW6*, pp. 294–299, 2007.

[17] K. Tokuda, T. Ogawa, K. Chiba, and S. Imai, "Spectral estimation of speech by mel-generalized cepstral analysis," *IEICE Transactions*, vol. 75, no. 7, pp. 1124–1134, 1992.