INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Low-dimensional representation of spectral envelope without deterioration for full-band speech analysis/synthesis system

Masanori Morise[1], *Genta Miyashita*[1], *Kenji Ozawa*[1]

[1]Faculty of Engineering, University of Yamanashi, Japan

mmorise@yamanashi.ac.jp

Abstract

A speech coding for a full-band speech analysis/synthesis system is described. In this work, full-band speech is defined as speech with a sampling frequency above 40 kHz, whose Nyquist frequency covers the audible frequency range. In prior works, speech coding has generally focused on the narrow-band speech with a sampling frequency below 16 kHz. On the other hand, statistical parametric speech synthesis currently uses the full-band speech, and low-dimensional representation of speech parameters is being used. The purpose of this study is to achieve speech coding without deterioration for full-band speech. We focus on a high-quality speech analysis/synthesis system and mel-cepstral analysis using frequency warping. In the frequency warping function, we directly use three auditory scales. We carried out a subjective evaluation using the WORLD vocoder and found that the optimum number of dimensions was around 50. The kind of frequency warping did not significantly affect the sound quality in the dimensions.
Index Terms: speech analysis/synthesis, speech coding, frequency warping, spectral envelope

1. Introduction

Statistical parametric speech synthesis (SPSS) [1] is being used all over the world and has recently been advanced by using deep neural networks (DNNs) [2]. DNNs require a high-performance speech analyzer to decompose speech into fundamental frequency (F0), spectral envelope, and aperiodicity by using the vocoder [3]. Merlin [4], a toolkit for building DNN models, utilized STRAIGHT [5] and WORLD [6]. Other parameters have also been proposed in GlottDNN [7]. WaveNet [8], a deep neural network for generating raw audio waveforms, can directly model the raw waveform of speech, but it requires an F0 for training. In automatic speech recognition (ASR), an acoustic modeling using the waveform has been proposed [9]. The difference between ASR and high-quality speech synthesis is the sampling frequency of the speech: ASR mainly uses the narrow-band speech with a sampling frequency below 16 kHz, while high-quality speech synthesis uses the full-band speech. It is difficult for processing using full-band speech to directly use the waveform. Therefore, parametric representation of speech is still important for SPSS.

SPSS has a problem in that it lacks the spectral-fine structure of speech. To compensate for this, a low-dimensional spectral feature extraction [10] has been proposed, where a deep auto-encoder to obtain the low-dimensional spectral parameter is used. The input is the power spectrum calculated from fast Fourier transform (FFT), and higher-quality speech can be synthesized compared with the spectral envelope estimated by high-quality vocoders.

Since the performance of the acoustic feature extraction by the auto-encoder depends on the training data, we focus on the low-dimensional representation of the spectral envelope by signal processing based on mel-cepstral analysis. The purpose of this study is to clarify the following two points for full-band speech.

- Number of appropriate dimensions for synthesizing speech as naturally as speech synthesized without coding.
- Effect of the frequency warping function on the sound quality in mel-cepstral analysis.

This knowledge will support DNN-based speech synthesis because the number of appropriate dimensions is used not only as the SPSS but also as the baseline for extracting the low-dimensional representation.

In Section 2 of this paper, we discuss related works on speech coding and give an overview of the subjective evaluation. In Section 3, we present the experimental conditions and give the results. In Section 4, we discuss the results and clarify the number of appropriate dimensions and the frequency warping function. We conclude in Section 5 with a brief summary and a mention of future works.

2. Speech coding in spectral envelope

There have been several advances in speech coding research for narrow-band speech. Linear predictive coding (LPC) [11] is one of the major algorithms, and line spectral pairs (LSP) [12] has been widely used in telecommunication systems. Cepstrum [13] is also a fundamental algorithm and is used for several improved algorithms. First, generalized cepstral analysis [14] was proposed, and mel-cepstral analysis [15, 16] followed a few years later.

Mel-generalized cepstral analysis [17] is widely used in speech synthesis research. It has a parameter for frequency warping, and users can optimize the parameter for narrow-band speech. In SPSS for full-band speech, there has been research on warped linear prediction [18]. GlottDNN [7] has also been proposed.

In mel-cepstral analysis, the mel-log spectrum approximation (MLSA) filter [19] can directly synthesize the speech waveform from the mel-cepstrum. In contrast, in vocoder-based synthesis, we can synthesize the speech from the decoded spectral envelope. We therefore propose directly using the frequency warping function on the basis of the auditory scales. The purpose of coding is to obtain the spectral envelope without deterioration for a full-band speech analysis/synthesis system.

2.1. Mel-cepstral analysis revisited

Figure 1 shows the outline of the proposed spectral coding and its parameters. In the frequency warping, there are three warping functions based on the auditory scales.

http://dx.doi.org/10.21437/Interspeech.2017-67

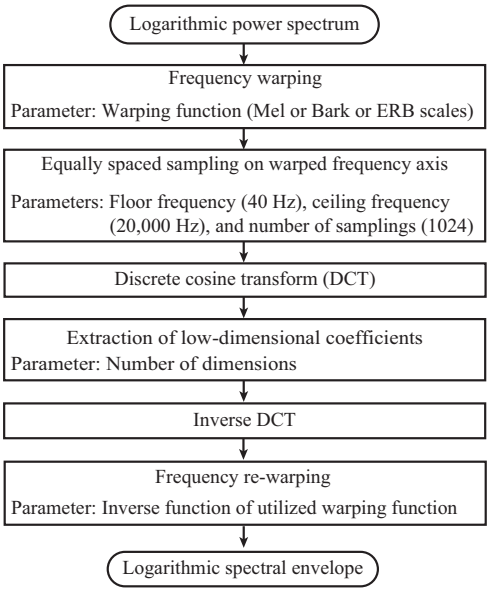

Figure 1: *Outline of proposed spectral coding. The values noted in brackets are the parameters used in this paper.*

The mel scale [20] is one of the most popular scales that is a perceptual scale of pitches. A popular function [21] is given by

$$m(f) = 1127.01048 \log\left(\frac{f}{700} + 1\right), \tag{1}$$

where f represents the input frequency in Hertz. This function has two coefficients, and several different coefficients have been proposed. We utilize the popular one.

Bark scale [22] is a psychoacoustical scale related to subjective measurements of loudness, and the function is given by

$$b(f) = 13 \arctan\left(0.00076f\right) + 3.5 \arctan\left(\left(\frac{f}{7500}\right)^2\right). \tag{2}$$

Since it is difficult to calculate its inverse function, it is approximately calculated by linear interpolation from $b(f)$ with a fine resolution.

Equivalent rectangular bandwidth (ERB) scale [23] is also a psychoacoustical scale, and it gives an approximation to the bandwidths of the filters in human auditory system. The function is given by

$$e(f) = 21.4 \log\left(\frac{4.37f}{1000} + 1\right). \tag{3}$$

The three warping functions are similar to the logarithmic warping, but the major difference is in the low frequency band. Figure 2 shows the difference among them. The vertical axis is normalized to indicate 1.0 at the frequency of 20 kHz. The warping functions determine the distribution of resolution from low to high frequencies.

2.2. Equally spaced sampling on the warped frequency axis

The warped spectrum is sampled at equal intervals on the warped frequency axis. This step has three parameters: the floor and ceiling frequencies and the number of samplings. We set the floor frequency to 40 Hz in accordance with the floor frequency

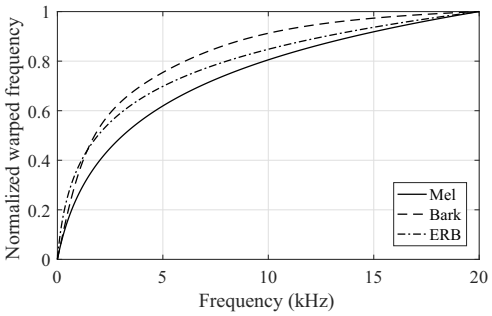

Figure 2: *Three warping functions based on each auditory scale. The vertical axis is normalized to indicate 1.0 at the frequency of 20 kHz.*

of general F0 estimation and set the ceiling frequency to 20,000 Hz, which is the upper limit of the audible frequency range of human beings.

The number of samplings is related to the maximum number of mel-cepstrum demensions and affects the accuracy of the decoded spectral envelope. We use WORLD, a high-quality speech analysis/synthesis system, with the FFT size of 2048 for the full-band speech. We know from sampling theory that the significant value is 1025, so we used 1024, which is close to this value. The values at frequencies used in sampling were calculated by simple linear interpolation.

2.3. Low-dimensional representation of the spectral envelope

The sampled sequence is transformed by discrete cosine transform (DCT), and then the liftering for extracting the low-dimensional coefficients is carried out. The number of dimensions for extraction is the parameter and determines the coding efficiency. When the dimension is set to N, the coefficients from 0 to N-1 are extracted, and the spectral envelope is compressed to N/1025. The extracted coefficients are transformed by inverse DCT to the logarithmic spectral envelope on the warped frequency axis. The warped spectral envelope is rewarped to the linear frequency axis by the inverse function of the warping function.

2.4. Input power spectrum

The input of mel-cepstral analysis has generally been the logarithmic power spectrum calculated by FFT. However, the calculated power spectrum depends on the temporal positions [24] even if the spectral envelope is temporally invariant. In a speech analysis/synthesis system, this time-varying component causes the sound quality to deteriorate. A temporally static representation of the spectral envelope has been proposed to overcome this problem.

STRAIGHT [5] is one of the most popular vocoders that has an algorithm for obtaining the temporally static spectral envelope. It uses the compensatory time window to remove the time-varying component. TANDEM-STRAIGHT [24, 25] and WORLD use the TANDEM window and CheapTrick [26, 27], respectively. Other algorithms such as F0-adaptive multi-frame integration analysis [28] have also been proposed to remove the time-varying component. In SPSS, the low-dimensional parameter [10] is extracted from the power spectrum by the deep autoencoder. This study showed that the power-spectrum-based

representation achieved the best sound quality compared with the same representation extracted from spectral envelopes estimated by STRAIGHT and WORLD. On the other hand, since the simple power spectrum cannot synthesize natural speech, the spectral envelope is still used for speech analysis/synthesis systems.

3. Subjective evaluation

A subjective evaluation was carried out to determine the number of appropriate dimensions in spectral envelope representation without deterioration compared with the spectral envelope without coding. The difference among frequency warping functions is also shown from the same evaluation. The subjective evaluation uses original and re-synthesized speech.

3.1. Vocoder used for the evaluation

We used WORLD (D4C edition [29]) as the high-quality vocoder utilized in the Merlin toolkit [4]. In each speech parameter estimation, we used DIO [30], CheapTrick [26, 27] and D4C [29] to measure F0, spectral envelope, and aperiodicity, respectively. In the synthesis from speech parameters, we did not use the MLSA filter but rather the default function in WORLD.

The F0 floor and ceiling frequencies in the F0 estimation were set to 71 and 800 Hz as defaults. The frame shift was set to 5 ms. We visually checked the estimation result in F0 and revised the definitive error in which a voiced segment was wrongly identified as an unvoiced segment. We also checked that the sound quality of the re-synthesized speech did not contain any fatal errors. The FFT size used in the spectral envelope was set to 2048 samples. There are several parameters in each estimator, but we did not change any of them from their defaults.

3.2. Conditions

Table 1 represents the conditions in the subjective evaluation. A multiple stimuli with hidden reference and anchor (MUSHRA) defined by ITU-R recommendation BS.1534-3 was used for the subjective evaluation. In the MUSHRA evaluation, the participants scored the speech stimuli on a scale of 0 to 100 (full marks) using a graphical user interface. In cases where there is no significant difference between a pair, it means that we cannot identify the difference by this test. MUSHRA can generally evaluate smaller differences than the mean opinion score (MOS). Non-significant (n.s.) does not guarantee that they have the same sound quality but rather suggests that we cannot detect the difference by both MUSHRA and MOS tests. Since these tests have generally been used to identify differences of sound quality, this evaluation would be enough to show the effectiveness of the coding. A sound-proof room with the A-weighted SPL of 18 dB was used, and 12 persons with normal hearing abilities participated in the evaluation. The sound stimuli were reproduced through a set of headphones (SENNHEISER HD650).

The speech stimuli used for the subjective evaluation were 20 words spoken by two men and two women. The sampling frequency and quantization bit were 48 kHz and 16 bits, respectively. A speech stimulus consisted of Japanese four-mora words including consonants.

The participants evaluated 14 speech stimuli at the same time. These 14 speech stimuli consisted of 12 speech stimuli synthesized with the coded spectral envelope and two speech stimuli (the original speech and speech synthesized by WORLD without coding). The 12 speech stimuli consisted of a product

Table 1: *Conditions of the subjective evaluation.*

Evaluation protocol	
Method	MUSHRA
Number of participants	12 persons
Environment	18 dB (A-weighted SPL)
Headphones	SENNHEISER HD650
Audio I/O	Roland QUAD-CAPTURE
Characteristics of the speech used in the evaluation	
Number of speakers	4 (2 males and 2 females)
Number of speech stimuli	20 (5 words per speaker)
Kind of speech	4-mora word
Sampling / Quantization	48 kHz / 16 bit
Conditions of the frequency warping	
Kinds of auditory scale	Mel, Bark, and ERB
Number of dimensions	20, 30, 40, and 50 dimensions
Floor/ceiling frequency	40/20,000 Hz
Number of samplings	1024

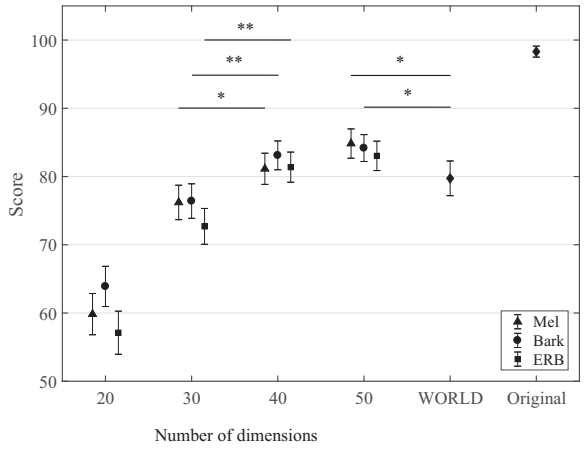

Figure 3: *Results of MUSHRA subjective evaluation. A single asterisk represents significant differences (Adjusted $p < 0.05$); double asterisks represent highly significant differences (Adjusted $p < 0.01$). To reduce the number of multiple comparisons, we limit the combinations required for the discussion.*

of three frequency-warping functions (the mel, Bark and ERB) and four dimensions (20, 30, 40 and 50 dimensions). These dimensions were determined by an exploratory listening test. In the evaluation, 14 speech stimuli were randomized.

3.3. Results

Figure 3 shows the experimental results. The vertical axis represents the average scores under each condition. The error bar represents a 95% confidence interval. We carried out a statistical analysis for the results. Since multiple testing is required for the discussion, the two-stage linear step-up procedure [31] based on the Benjamini and Hochberg procedure [32] was carried out.

Table 2 shows the list for multiple comparisons. Values noted in brackets represent the number of dimensions used in each frequency warping function. In cases where the adjusted p-value exceeds the reference value, $n.s.$ (non-significant) is shown instead of the adjusted p-value. The difference between 30 and 40 was significant in all auditory scales, while the difference between 40 and 50 was not. In the comparison among auditory scales in the same dimension, there was no significant

Table 2: *The list of multiple comparisons and their adjusted p-values. Values noted in brackets represent the number of dimensions. In the case where the adjusted p-value exceeds the reference value, it is shown as $n.s.$ (non-significant).*

Combination	Adjusted p-value [33]
WORLD and mel (50)	0.020
WORLD and Bark (50)	0.030
WORLD and ERB (50)	n.s.
Mel (40) and Mel (50)	n.s.
Bark (40) and Bark (50)	n.s.
ERB (40) and ERB (50)	n.s.
Mel (30) and Mel (40)	0.0206
Bark (30) and Bark (40)	0.0008
ERB (30) and ERB (40)	0.00002
Mel (30) and Bark (30)	n.s.
Mel (30) and ERB (30)	n.s.
Bark (30) and ERB (30)	n.s.
Mel (40) and Bark (40)	n.s.
Mel (40) and ERB (40)	n.s.
Bark (40) and ERB (40)	n.s.
Mel (50) and Bark (50)	n.s.
Mel (50) and ERB (50)	n.s.
Bark (50) and ERB (50)	n.s.

difference in 30, 40, and 50 dimensions. In the comparison between WORLD and low-dimensional representation with 50 dimensions, there were significant differences in the mel and Bark scales, but the low-dimensional representations were superior to WORLD in the sound quality. The reason for this is discussed in the next section.

The results provide us with two main findings.

- The number of appropriate dimensions was around 50.
- There was no significant difference among auditory scales in the case where 50 dimensions were used.

In the following discussion, we analyze and explain why the low-dimensional representation was superior to WORLD.

4. Discussion

We first explain why some of the low-dimensional representations was superior to the spectral envelope without coding. After that, we discuss the coding efficiency compared with the conventional approach.

4.1. Improvement of sound quality by coding

The results showed that the mel and Bark scales in 50 dimensions were significantly superior to those of WORLD. Figure 4 shows the results for each speaker (mel scale (50 dimensions) and WORLD only). There were significant differences in male 2 and female 2.

Qualitative analysis of speech stimuli showed that the temporal difference between neighboring frames tends to be larger compared with others. The voiced sound often contains breath, and breath can cause decay of the SNR. It also causes error in the estimated parameters. In particular, since the spectral envelope and aperiodicity estimations use the F0 information, error with F0 affects the whole accuracy. It seems that the low-dimensional representation can remove the temporal difference caused by the F0 error and obtain the robust spectral envelope. Our future work will involve carrying out another evaluation to verify this hypothesis.

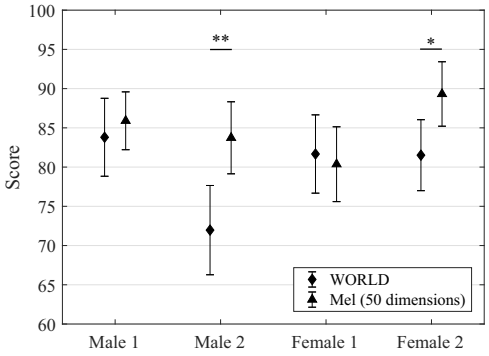

Figure 4: *Comparison between WORLD and mel scale (50 dimensions) for each speaker. A significant difference in two speakers is evident.*

4.2. Comparison of coding efficiency compared with conventional approach

The low-dimensional representation by the deep auto-encoder [10] can compress the spectral envelope to 59 dimensions for full-band speech. In this research, since the F0 and aperiodicity were represented by one and 25 dimensions, the number of dimensions is 85 dimensions per frame. GlottDNN [7] used 111 dimensions per frame. In contrast, we can achieve a better compression with only signal processing without any statistical technique that depends on the training data. We have also shown that the aperiodicity can be compressed to just five dimensions [29]. In the case where the dimensions for the spectral envelope representation were set to 50, we can compress the speech parameters to 56 dimensions (one for F0, 50 for spectral envelope, and five for aperiodicity) per frame. Additional compression would be possible by using an auto-encoder with the compressed parameters.

5. Conclusion

In this paper, we revisited mel-cepstral analysis, and directly used the frequency warping function based on auditory scales. Subjective evaluation using full-band speech was carried out to show the number of appropriate dimensions for representing the spectral envelope and the types of warping function. The results showed that 50 dimensions were enough to synthesize the speech as naturally as the speech synthesized by WORLD without coding. Furthermore, the types of warping function did not significantly affect the sound quality in the case where the number of dimensions was 50.

The next step is the quantization coding in speech parameters. To show the appropriate frame shift is also an important work. A 5-ms analysis seems to be enough for the speech with F0 below 200 Hz because the fundamental period is 5 ms. However, since the female speech often has an F0 above 200 Hz, a small frame shift below 5 ms may improve the sound quality. These optimized parameters would be helpful for not only speech analysis/synthesis systems but also SPSS research using full-band speech.

6. Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP16H01734, JP15H02726, JP16H05899, JP16K12511.

7. References

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, pp. 1039–1064, 2009.

[2] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," *in Proc. ICASSP2013*, pp. 7962–7966, 2013.

[3] H. Dudley, "Remaking speech," *J. Acoust. Soc. Am.*, vol. 11, no. 2, pp. 169–177, 1939.

[4] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," *in Proc. of SSW 2016*, pp. 218–223, 2016.

[5] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based F0 extraction," *Speech Communication*, vol. 27, no. 3–4, pp. 187–207, 1999.

[6] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. & Syst.*, vol. E99-D, pp. 1877–1884, 2016.

[7] M. Airaksinen, B. Bollepalli, L. Juvela, Z. Wu, S. King, and P. Alku, "Glottdnn — a full-band glottal vocoder for statistical parametric speech synthesis," *in Proc. INTERSPEECH2016*, pp. 2473–2477, 2016.

[8] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[9] Z. Tüske, P. Golik, R. Schlüter, and H. Ney, "Acoustic modeling with dpee neural networks using raw time signal for LVCSR," *in Proc. INTERSPEECH 2014*, pp. 890–894, 2014.

[10] S. Takaki and J. Yamagishi, "A deep auto-encoder based low-dimensional feature extraction from FFT spectral envelopes for statistical parametric speech synthesis," *in Proc. ICASSP2016*, pp. 5535–5539, 2016.

[11] B. Atal and S. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Am.*, vol. 50, no. 2B, pp. 296–302, 1971.

[12] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signals," *J. Acoust. Soc. Am.*, vol. 57, no. S1, p. S35, 1975.

[13] A. Oppenheim and R. Schafer, "Homomorphic analysis of speech," *IEEE Trans. Audio and Electroacoust.*, vol. AU-16, no. 2, pp. 221–226, 1968.

[14] K. Tokuda, T. Kobayashi, and S. Imai, "Generalized cepstral analysis of speech —unified approach to LPC and cepstral method," *in Proc. ICSLP-90*, pp. 37–40, 1990.

[15] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," *in Proc. ICASSP92*, vol. 1, pp. 137–140, 1992.

[16] K. Tokuda, "Speech coding based on adaptive melcepstral analysis," *in Proc. ICASSP'94*, pp. 197–200, 1994.

[17] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analyasis — a unified approach to speech spectral estimation," *in Proc. ICSLP-94*, pp. 1043–1046, 1994.

[18] T. Raitio, A. Suni, M. Vainio, and P. Alku, "Wideband parametric speech synthesis using warped linear prediction," *in Proc. INTERSPEECH2012*, pp. 1420–1423, 2012.

[19] S. Imai, "Mel log spectrum approximation (MLSA) filter for speech synthesis," *Electronics and Communications in Japan (Part I Communications)*, vol. 66, no. 2, pp. 10–18, 1983.

[20] S. Stevens, J. Volkmann, and E. Newman, "A scale for the measurement of the psychological magnitude pitch," *J. Acoust. Soc. Am.*, vol. 8, no. 3, pp. 185–190, 1937.

[21] S. Stevens and J. Volkmann, "The relation of pitch to frequency: A revised scale," *The American Journal of Psychology*, vol. 53, no. 3, pp. 329–353, 1940.

[22] E. Zwicker and H. Fastl, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," *J. Acoust. Soc. Am.*, vol. 68, no. 5, pp. 1523–1525, 1980.

[23] B. Moore, *Psychology of Hearing*. Academic Press, 2003.

[24] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation," *in Proc. ICASSP2008*, pp. 3933–3936, 2008.

[25] H. Kawahara and M. Morise, "Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework," *SADHANA - Academy Proceedings in Engineering Sciences*, vol. 36, no. 5, pp. 713–728, 2011.

[26] M. Morise, "CheapTrick, a spectral envelope estimator for high-quality speech synthesis," *Speech Communication*, vol. 67, pp. 1–7, 2015.

[27] ——, "Error evaluation of an f0-adaptive spectral envelope estimator in robustness against the additive noise and f0 error," *IEICE Trans. Inf. & Syst.*, vol. E98-D, no. 7, pp. 1405–1408, 2015.

[28] T. Nakano and M. Goto, "A spectral envelope estimation method based on f0-adaptive multi-frame integration analysis," *in Proc. SAPA-SCALE 2012*, pp. 11–16, 2012.

[29] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57–65, 2016.

[30] M. Morise, H. Kawahara, and H. Katayose, "Fast and reliable f0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech," *in Proc. AES 35th International Conference, CD-ROM*, pp. CD–ROM, 2009.

[31] Y. Benjamini, A. M. Krieger, and D. Yekutieli, "Adaptive linear step-up procedures that control the false discovery rate," *Biometrika*, vol. 93, no. 3, pp. 491–507, 2006.

[32] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J. R. Statist. Soc. B*, vol. 57, no. 1, pp. 289–300, 1995.

[33] S. P. Wright, "Adjusted p-values for simultaneous interface," *Biometrics*, vol. 48, pp. 1005–1013, 1992.