EFFECT OF MPEG AUDIO COMPRESSION ON VOCODERS USED IN STATISTICAL PARAMETRIC SPEECH SYNTHESIS

Bajibabu Bollepalli[*], *Tuomo Raito*[†]

[*] Department of Speech, Music and Hearing, KTH, Stockholm, Sweden
[†] Department of Signal Processing and Acoustics, Aalto University, Finland

ABSTRACT

This paper investigates the effect of MPEG audio compression on HMM-based speech synthesis using two state-of-the-art vocoders. Speech signals are first encoded with various compression rates and analyzed using the GlottHMM and STRAIGHT vocoders. Objective evaluation results show that the parameters of both vocoders gradually degrade with increasing compression rates, but with a clear increase in degradation with bit-rates of 32 kbit/s or less. Experiments with HMM-based synthesis with the two vocoders show that the degradation in quality is already perceptible with bit-rates of 32 kbit/s and both vocoders show similar trend in degradation with respect to compression ratio. The most perceptible artefacts induced by the compression are spectral distortion and reduced bandwidth, while prosody is better preserved.

Index Terms— Statistical parametric speech synthesis, HMM, MPEG, MP3, GlottHMM, STRAIGHT

1. INTRODUCTION

Research on text-to-speech (TTS) synthesis has taken steps from read-aloud corpus based synthesis of short sentences to audio-book based synthesis of longer paragraphs [1]. Nowadays, one can find extensive amounts of speech data from, e.g., the world wide web. However, due to the limitations in storage and bandwidth, speech data is typically available in compressed forms. In addition, speech data are expressed in various forms involving also mixtures of speech, music and video. Thus, instead of using speech-specific compression methods, general audio compression methods are often used when speech data is distributed on the Internet. Depending on the optimization of video and audio data rate, compression may introduce severe artefacts in the speech signal.

There are a few studies that have addressed the degradation of speech parameters due to compression (see e.g. [2,3]). In [4], the authors of the current paper conducted the first study on how the compression of speech affects vocoding and statistical parametric speech synthesis. The results of the study indicated that building voices from compressed speech data was not severely affected if the compression rate was 32 kbit/s or more. In this paper, the previous study is elaborated by including two different vocoding techniques and using a more detailed subjective evaluation. Different feature extraction algorithms in the two vocoders are expected to behave differently in relation to speech compression, and the different data representation in statistical modeling and synthesis technique may also affect the quality of synthesized speech. Moreover, listening tests are conducted with only synthetic speech in order to more accurately study the effect of compression, while in previous study, vocoded and natural speech were included in the same listening test.

The paper is structured as follows. In Section 2, speech compression using MPEG-1 Audio Layer III (MP3) audio compression techniques is described. Section 3 describes the two vocoders, GlottHMM and STRAIGHT, used in this study. In Section 4, the effect of compression at different bit-rates on the vocoder parameters is first studied using objective methods, after which the role of speech compression in the quality of HMM-based synthesis is studied using subjective listening tests. Finally, Section 5 discusses the obtained results and summarizes the findings of the paper.

2. SPEECH COMPRESSION

MPEG-1 Audio Layer 3 compression method [5], commonly known as MP3 was used for compressing speech in this study. MPEG (moving pictures expert group) is a standard in audio

Table 1. Bit-rates and corresponding theoretical and realized compression ratios with respect to 256 kbit/s 16 kHz PCM speech.

Bit-rate (kbit/s)	Compression ratio w.r.t. bit-rate	Compression ratio w.r.t. file size
160	1.6	1.56
128	2	1.92
64	4	3.13
32	8	6.25
24	10.67	8.33
16	16	12.50
8	32	25.00

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287678 (Simple[4]All).

coding which enables high compression rates while preserving high quality. MP3 takes advantage of the characteristics of human auditory mechanism to compress audio. MP3 compression is lossy; it uses psychoacoustic models to reduce the precision of components less audible to human hearing, and encodes the remaining material with high efficiency. In MPEG compression, the audio signal is first converted into spectral components using a filter bank analysis. For each spectral component, the perceptual masking effect caused by other components is first calculated. Then, each spectral component is quantized so that the low-level signals (maskee) can be coded with fewer bits than the simultaneous occurring stronger signal (masker) as long as the masker and maskee are close enough to each other in frequency and time [6], thus keeping the quantization noise below the masking threshold. With very low bit-rates, low-pass filtering is used in order to reduce audio bandwidth and thus the required bit-rate.

In this work, a freely available software called the *LAME-v3.99* [7] encoder is used to compress speech signals with standard options (fixed bit-rate encoding scheme). Table 1 shows the bit-rates along with the compression ratios used in this study. Here, compression ratios are calculated with respect to the original speech utterances recorded at a sampling rate of 16 kHz with 16-bit resolution, resulting in a data rate of 256 kbit/s with pulse code modulation (PCM) encoding.

3. VOCODERS

3.1. GlottHMM

GlottHMM [8, 9] is designed for parameter extraction and speech waveform generation for HMM-based speech synthesis. GlottHMM aims to accurately model the speech production mechanism by using glottal inverse filtering. GlottHMM has been shown to yield high-quality synthetic speech [8–12], better or comparable to the quality of STRAIGHT [13], the most widely used vocoder in HMM-based speech synthesis.

In GlottHMM speech parametrization, iterative adaptive inverse filtering (IAIF) [14] is used to estimate the vocal tract filter and the voice source signal. Linear prediction (LP) is used for spectral estimation in the IAIF method, and the estimated vocal tract filter is converted to line spectral frequencies (LSF) [15] for better representation of the LP information in HMM-training. From the estimated voice source signal, fundamental frequency (F0) is estimated with the autocorrelation method, and the log harmonic-to-noise ratio (HNR) of five frequency bands is estimated by comparing the upper and lower spectral envelopes constructed from the harmonic peaks and the interharmonic valleys, respectively. HNR values are then averaged to five frequency bands according to the equivalent rectangular bandwidth [16] (ERB) scale. Additionally, the voice source spectrum is estimated with LP (converted to LSFs) in order to control the phonation characteristics in synthesis. The GlottHMM parameters are shown in Table 2.

In synthesis, a pre-stored natural glottal flow pulse is used for reconstructing the excitation signal. The pulse is first interpolated to a duration according to F0 and scaled in amplitude according to the energy parameter. In order to match the degree of voicing in the excitation, noise is added according to the HNR of five bands in the spectral domain. In order to control the phonation type, the excitation spectrum is matched to the given voice source LP spectrum. Finally, the excitation is filtered with the vocal tract filter to synthesize speech.

Table 2. Speech features for the GlottHMM vocoder.

GlottHMM features	Number of parameters
Vocal tract spectrum	30
Voice source spectrum	10
Harmonic-to-noise ratio (HNR)	5
Energy	1
Fundamental frequency (F0)	1

3.2. STRAIGHT

STRAIGHT [13, 17] was originally proposed as a speech manipulation tool, but nowadays it is widely used for HMM-synthesis [18]. STRAIGHT extracts three types of parameters: F0, spectrum, and aperiodicity parameters (AP). In STRAIGHT analysis, F0 and voiced-unvoiced decision are first estimated using an instantaneous-frequency based algorithm and a fixed-point analysis TEMPO [19]. In order to estimate speech spectrum, F0-adaptive smoothing is applied to remove the effect of signal periodicity, after which filter coefficient are estimated with mel-cepstrum (MCEP) [20]. The AP for mixed excitation are based on an amplitude ratio between the lower and upper smoothed spectral envelopes [17] and averaged across 21 frequency sub-bands. The STRAIGHT parameters are shown in Table 3.

STRAIGHT synthesis uses mixed excitation [21] consisting of impulses and a noise component weighted according to the AP. The pitch-synchronous overlap add (PSOLA) [22] method is used to reconstruct the excitation signal, which excites a mel log spectrum approximation (MLSA) filter [23].

4. EXPERIMENTS

4.1. Speech material

Two databases designed for TTS development were used in experiments. The first corpus consists of 599 sentences by a Finnish male (labeled as MV), and the second one consists of 513 sentences by a Finnish female (labeled as HK). All

Table 3. Speech features for the STRAIGHT vocoder.

STRAIGHT features	Number of parameters
Mel-cepstrum	40
Aperiodicity parameters (AP)	21
Fundamental frequency (F0)	1

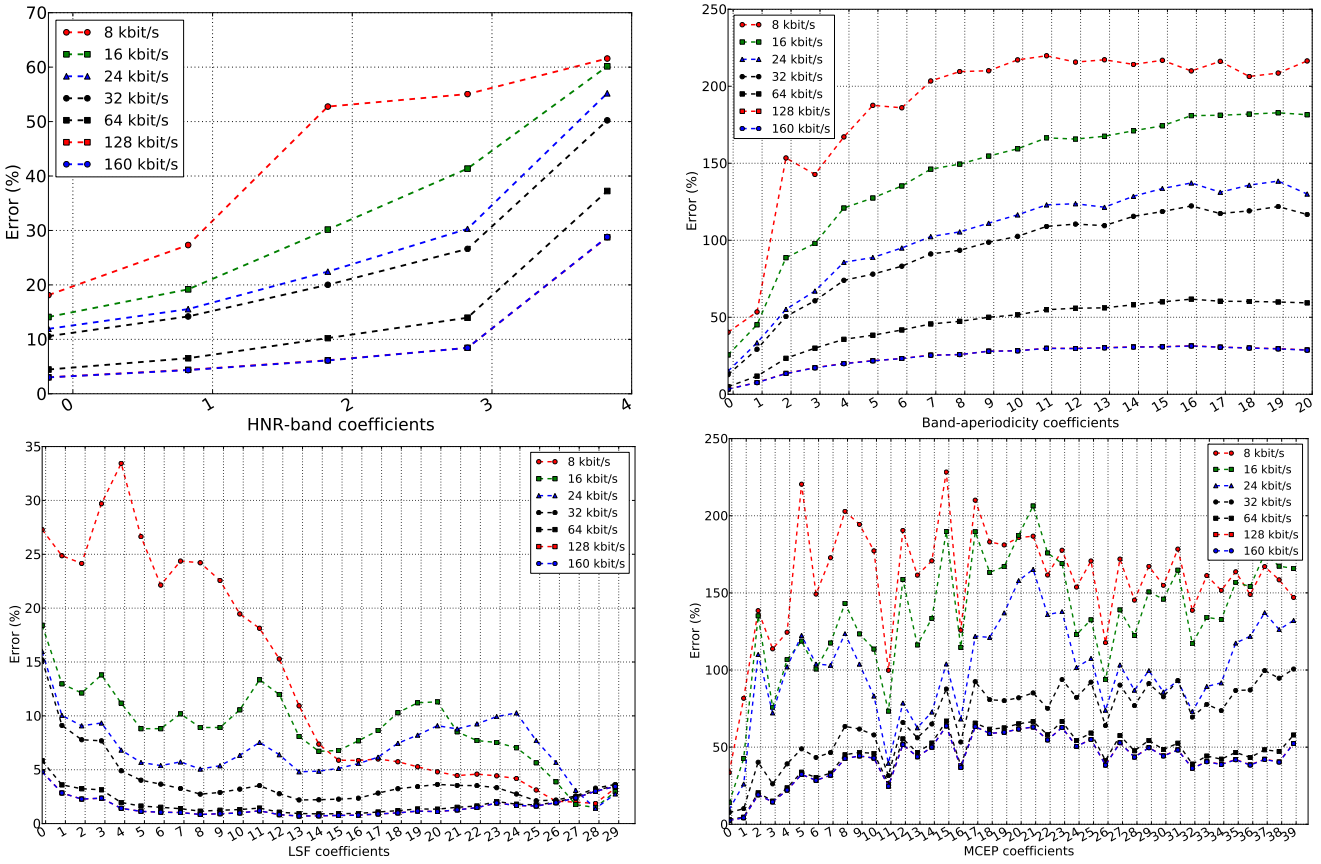

Fig. 1. Relative error of HNR and LSF (GlottHMM, left) and AP and MCEP (STRAIGHT, right) as a function of bit-rate.

audio files were PCM encoded and sampled at 16 kHz with a resolution of 16 bits, resulting in a data rate of 256 kbit/s.

4.2. Objective evaluations of vocoder parameters

The effect of compression was evaluated by comparing the vocoder parameters extracted from the MP3-processed sounds to those obtained from the uncompressed ones. For each compression rate, the relative error was determined between the parameter values computed from the uncompressed and compressed sound for both speakers. Note that the relative error depends on the scale of the original parameter values. However, relative error was used in order to have a common error measure for all the parameters of the two vocoders, and because it seems to describe the effects of compression fairly well. The following three types of parameters were analyzed: 1) F0, 2) HNR/AP, 3) LSF/MCEP.

Compression has only a small effect on F0 error, stemming mainly from slight differences in estimated F0 values and voicing decisions. The relative error of F0 is 1–5 % for GlottHMM and 1–3 % for STRAIGHT, but the differences between bit-rates were not statistically significant across the two voices. However, GlottHMM seems to be slightly more affected by the compression than STRAIGHT with low bit-rates. Figure 1 shows the relative error of HNR and LSF for GlottHMM and AP and MCEP for STRAIGHT. For both vocoders, the error of HNR/AP is rather small with high bit-rates such as 160 kbit/s and 128 kbit/s, with only a small increase in error with 64 kbit/s. With bit-rates 32 kbit/s and lower, however, the error increases substantially. The relative error of LSF for GlottHMM shows a similar effect: high bit-rates (64 kbit/s or more) show small errors, while 32 kbit/s and lower bit-rates show larger errors. Particularly the 8 kbit/s voice shows very high errors, especially in the perceptually important low and mid-frequencies. MCEP parameters of STRAIGHT also show that high bit-rates ($\geq$ 64 kbit/s) show small errors while higher compression significantly affects the spectral parameters. In conclusion, although the degradations are gradual, the compressed acoustic signals are significantly different from the original signal if the bit-rate is 32 kbit/s or less, which has a clear effects on both vocoders.

Experiments using different LSF (LP analysis) and MCEP orders were also conducted. LSF order was varied from 14 to 30 and MCEP order from 10 to 40. The results indicate that increasing the LSF order had an effect of reducing the average parameter error, whereas increasing the MCEP order had the opposite effect of increasing the error due to compression.

4.3. Evaluation of HMM synthesis quality

HMM-based synthetic voices were built with both vocoders and voices and five bit-rates. Standard HTS procedure [24,25]

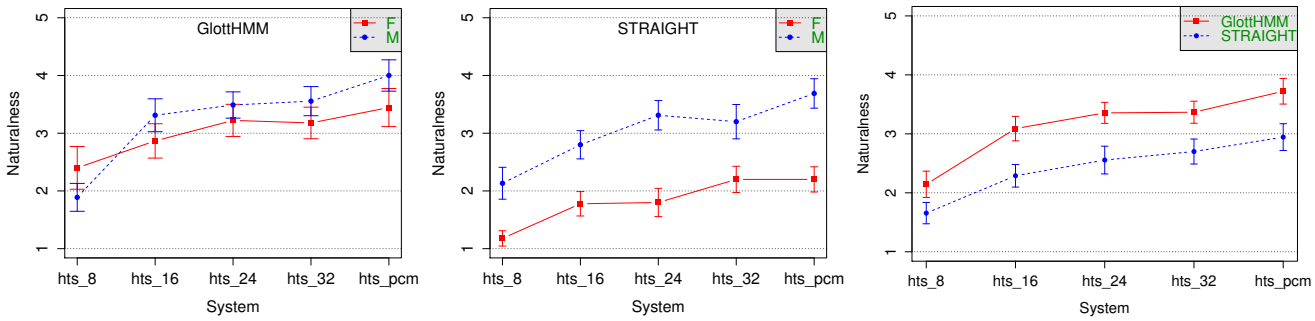

Fig. 2. HMM-based synthesis naturalness scores as a function of bit-rate for GlottHMM (leftmost figure) and STRAIGHT (center figure) for the female (F) and male (M) speakers, and averaged MOS scores for both vocoders (rightmost figure).

was used for training the voices. Subjective evaluations were conducted to assess the quality of the HMM-based voices. As subjective evaluations are more laborious than objective ones, only five bit-rates were included in the tests, concentrating on the low bit-rates, where the differences are expected to be perceptible [4]: 1) Full PCM (256 kbit/s), 2) 32 kbit/s, 3) 24 kbit/s, 3) 16 kbit/s, and 4) 8 kbit/s. For each voice (5 bit-rates, 2 vocoders, 2 genders), 3 randomly selected sentences were used, totaling 60 sentences per test. The sentences were presented in random order to subjects (15 listeners, of which 11 native Finnish) who rated the naturalness of signals on the mean opinion score (MOS) scale, ranging from 1 to 5 (1–completely unnatural, 5–completely natural).

Figure 2 shows the means and 95% confidence intervals of the MOS ratings for each vocoder and speaker, and also averages across gender. For GlottHMM, male and female voices are rated similar in general with a slight preference for the male voice except with the lowest bit-rate. The low quality of the male 8 kbit/s voice seems to stem from the overly sharp formants in the mid-frequencies, which can be seen in Figure 3 as over-emphasized frequencies from 4 kHz to 5 kHz. The female voice, however, exhibits rather loss of mid-frequencies at low bit-rates and is thus perceived overly soft, but not as low in quality as the male voice. For STRAIGHT, the male and female voices are rated completely different; STRAIGHT male voices are comparable to the GlottHMM voices, but female voice is rated very low. This may stem from the simpler mixed excitation used instead of the glottal-flow excitation of GlottHMM. Both male and female low bit-rate STRAIGHT voices exhibit overly sharp formants at mid-frequencies, which can be seen in Figure 3. On average, GlottHMM is rated always better than STRAIGHT, but the degradation due to compression seems to follow the same general trend; the quality gradually decreases as a function of bit-rate, which is perceptible with bit-rates of 32 kbit/s and lower (although not statistically significant).

Figure 3 shows the long-term average spectra of natural and synthetic speech with different bit-rates, plotted separately for each vocoder and speaker. The spectra of natural compressed speech shows that low-pass filtering is used in encoding. With 32 kbit/s, spectral components are missing above 7 kHz, and for lower bit-rates the cut-off frequency is between 5 kHz and 6 kHz. This introduces clear audible effects, and seems to affect the spectral estimation and modeling in the boundary frequencies in the synthetic voices. In addition, serious deviations in spectrum from 1 kHz to 5 kHz can be observed with the lowest bit-rates.

The most perceptible artefacts induced by the compression were spectral distortion due to the overly sharp spectral components at 4 kHz–5 kHz, especially with the two lowest bit-rates, and the reduced bandwidth induced by the compression with 24 kbit/s or lower bit-rates. However, the prosody of all voices was rather well preserved.

5. CONCLUSIONS

In this paper, the effects of using MP3-compressed speech in HMM-based speech synthesis was studied. Speech signals were encoded with various compression rates and experiments were performed using the GlottHMM and STRAIGHT vocoders. Objective evaluations showed that the parameters of both vocoders gradually degraded with increasing compression rates, but with a clear increase in degradation with bit-rates of 32 kbit/s or less. Experiments with HMM-based speech synthesis showed that the degradation of subjective quality was perceptible with bit-rates of 32 kbit/s or less, and both vocoders showed similar trend in degradation with respect to compression ratio. The most perceptible artefacts induced by the compression were spectral distortion and reduced bandwidth, while prosody was better preserved.

6. REFERENCES

[1] S. King and V. Karaiskos, "The Blizzard Challenge 2012," in *The Blizzard Challenge 2012 workshop*, 2012, http://festvox.org/blizzard.

[2] J. Gonzalez and T. Cervera, "The effect of MPEG audio compression on a multi-dimensional set of voice parameters," *Log. Phon. Vocol.*, vol. 26, no. 3, pp. 124–138, 2001.

[3] R.J.J.H. van Son, "A study of pitch, formant, and spectral estimation errors introduced by three lossy speech compression algorithms," *Acta Acustica United With Acustica*, vol. 91, no. 4, pp. 771–778, 2005.

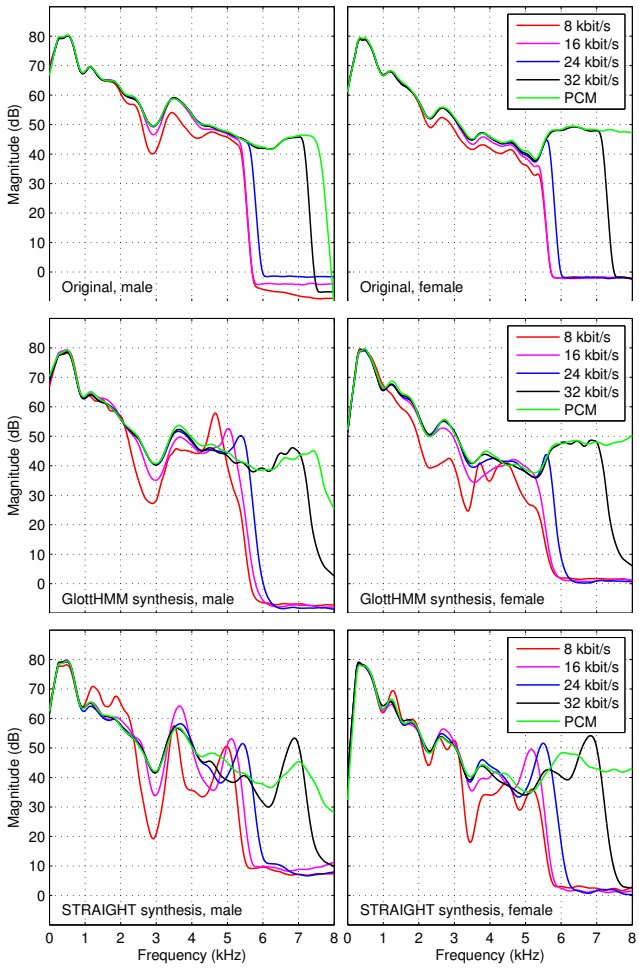

Fig. 3. Long-term average spectra of compressed and PCM speech (upper graphs) and synthetic speech with GlottHMM (middle graphs) and STRAIGHT (bottom graphs) with different bit-rates for the male (left) and female (right) speakers.

[4] B. Bollepalli, T. Raitio, and P. Alku, "Effect of MPEG audio compression on HMM-based speech synthesis," in *Proc. Interspeech*, 2013, pp. 1062–1066.

[5] ISO, "Information Technology – Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to About 1.5 Mbit/s – Part 3: Audio," 1993, ISO/IEC 11172-3:1993, International Organization for Standardization.

[6] G. Tzanetakis and P. Cook, "Sound analysis using MPEG compressed audio," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Proc.*, 2000, vol. 2, pp. 761–764.

[7] [Online], "LAME encoder," 2013, http://lame.sourceforge.net/.

[8] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Trans. Audio Speech Lang. Proc.*, vol. 19, no. 1, pp. 153–165, 2011.

[9] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Proc.*, 2011, pp. 4564–4567.

[10] A. Suni, T. Raitio, M. Vainio, and P. Alku, "The GlottHMM speech synthesis entry for Blizzard Challenge 2010," in *The Blizzard Challenge 2010 workshop*, 2010, http://festvox.org/blizzard.

[11] A. Suni, T. Raitio, M. Vainio, and P. Alku, "The GlottHMM entry for Blizzard Challenge 2011: Utilizing source unit selection in HMM-based speech synthesis for improved excitation generation," in *The Blizzard Challenge 2011 workshop*, 2011, http://festvox.org/blizzard.

[12] A. Suni, T. Raitio, M. Vainio, and P. Alku, "The GlottHMM entry for Blizzard Challenge 2012 – Hybrid approach," in *The Blizzard Challenge 2012 workshop*, 2012, http://festvox.org/blizzard.

[13] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, no. 3–4, pp. 187–207, 1999.

[14] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Commun.*, vol. 11, no. 2–3, pp. 109–118, 1992.

[15] F. K. Soong and B.-H. Juang, "Line spectrum pair (LSP) and speech data compression," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Proc.*, 1984, vol. 9, pp. 37–40.

[16] B. C. J. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *J. Acoust. Soc. Am.*, vol. 74, pp. 750–753, 1983.

[17] H. Kawahara, Jo Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," in *2nd Inernational Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, 2001.

[18] Heiga Zen, Keiichi Tokuda, and Alan W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, 2009.

[19] H. Kawahara, H. Katayose, A. de Cheveigné, and R. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity," in *Proc. Eurospeech*, 1999, pp. 2781–2784.

[20] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Proc.*, 1983, vol. 8, pp. 93–96.

[21] T. Yoshimura, K. Tokuda, T. Masuko, and T. Kitamura, "Mixed-excitation for HMM-based speech synthesis," *Proc. Eurospeech*, pp. 2259–2262, 2001.

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

[23] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Proc.*, 1992, vol. 1, pp. 137–140.

[24] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Sixth ISCA Workshop on Speech Synthesis*, 2007, pp. 294–299.

[25] [Online], "HMM-based speech synthesis system (HTS)," 2013, http://hts.sp.nitech.ac.jp.