Effect of MPEG Audio Compression on HMM-based Speech Synthesis

Bajibabu Bollepalli[1], *Tuomo Raitio*[2], *Paavo Alku*[2]

[1]Department of Speech, Music and Hearing, KTH, Stockholm, Sweden
[2]Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland
bajibabu@kth.se, tuomo.raitio@aalto.fi

Abstract

In this paper, the effect of MPEG audio compression on HMM-based speech synthesis is studied. Speech signals are encoded with various compression rates and analyzed using the GlottHMM vocoder. Objective evaluation results show that the vocoder parameters start to degrade from encoding with bit-rates of 32 kbit/s or less, which is also confirmed by the subjective evaluation of the vocoder analysis-synthesis quality. Experiments with HMM-based speech synthesis show that the subjective quality of a synthetic voice trained with 32 kbit/s speech is comparable to a voice trained with uncompressed speech, but lower bit rates induce clear degradation in quality.
Index Terms: speech synthesis, HMM, MP3, GlottHMM

1. Introduction

The paradigm of text-to-speech (TTS) has shifted 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 on the Internet, speech data is typically available in compressed forms. In addition, speech data are expressed in various forms which might involve also mixtures of voice signals, music and video. Thus, instead of using speech-specific compression methods, general audio compression methods are increasingly used when speech data is disseminated on the Internet. Depending on the optimization of video and audio data rate, compression may introduce severe artefacts in the speech signal.

In [2], it was shown that both 29 acoustical voice parameters, obtained from the multi-dimensional voice program (MDVP) model of Kay Elemetrics Corp., and the amplitude-frequency spectrum have high fidelity with high data rates (128 kbit/s and 168 kbit/s with sampling rate of 44.1 kHz) whereas lower bit-rates, such as 64 kbit/s or lower, introduce substantial modifications in the speech signal and amplitude spectrum parameters. In [3], it was shown that bit-rates of 80 kbit/s and up (with sampling rate of 44.1 kHz) can be used for acoustic analysis (pitch and formant extraction, global spectral measure, the spectral center of gravity) without any degradation in quality while a low bit-rate of 40 kbit/s introduces larger errors in formant measurements. However, to the best of our knowledge, there are no studies which specifically have addressed how compression of speech affects vocoding and statistical parametric speech synthesis [4].

The aim of this investigation is to study how compression of speech affects hidden Markov model (HMM) based speech synthesis. One of the most commonly used audio compression techniques, MPEG-1 Audio Layer III (MP3) [5], is used to compress speech signals with various bit-rates. All the analysis and synthesis steps are performed using the GlottHMM vocoder [6]. First, the extent to which the speech signal and the vocoder parameters are affected by the compression at different bit-rates is studied with objective methods. Second, the analysis-synthesis quality of the vocoder is evaluated by subjective listening tests by varying the bit-rate of the input speech. Finally, the role of speech compression in the quality of HMM-based synthesis is studied by building voices with various bit-rates and evaluating the subjective quality of the resulting synthetic voices.

2. Speech Compression

For compression of speech, we used the MPEG-1 Audio Layer 3 compression method [5], commonly known as MP3. MPEG (moving pictures expert group) is a standard in audio 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 discard or reduce precision of components less audible to human hearing, and encodes the remaining material with high efficiency. First, the audio signal is converted into spectral components using a filter bank analysis. For each spectral component, the perceptual masking effect caused by other components is calculated. Later, the low-level signals (maskee) are replaced by a simultaneous occurring stronger signal (masker) as long as the masker and maskee are close enough to each other in frequency or time [7].

In this work, we have used a freely available software called the *LAME-v3.99* [8] encoder to compress speech signals with different bit-rates. The standard options of the encoder were used, i.e., the fixed bit-rate encoding scheme. All manipulations were done on a PC workstation running Linux. 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 which are 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.

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

3. Vocoder

The GlottHMM statistical parametric speech synthesizer [6] is used in the experiments of this study. GlottHMM aims to accurately model the speech production mechanism by decomposing speech into the vocal tract filter and the voice source signal using glottal inverse filtering. It is built on a basic framework of a HMM-based speech synthesis system (HTS) [9, 10], but it uses a distinct type of vocoder for parameterizing and synthesizing speech. GlottHMM has been shown to yield high-quality synthetic speech [6, 11, 12, 13], better or comparable to the quality of STRAIGHT [14], which is currently the most commonly used vocoder in statistical parametric speech synthesis.

In the parametrization of speech with GlottHMM, iterative adaptive inverse filtering (IAIF) [15] 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) [16] for better representation of the LP information in HMM-training [17]. From the estimated voice source signal, fundamental frequency (F0) is estimated with the autocorrelation method, and the harmonic-to-noise ratio (HNR) is estimated in five bands according to the equivalent rectangular bandwidth (ERB) [18] scale: 0–241 Hz, 241–731 Hz, 731–1735 Hz, 1735–3791 Hz, and 3791–8000 Hz. The voice source spectrum is estimated with LP and converted to LSFs.

In synthesis, natural glottal flow pulses are used for reconstructing the excitation signal. Pulses are interpolated in time and scaled in amplitude to match F0 and energy. 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. Also, in order to control the phonation type, the excitation is filtered with an infinite impulse response (IIR) filter to match the voice source spectrum. Finally, the created excitation is filtered with the vocal tract filter to synthesize speech.

Besides being used as a key element of statistical speech synthesis, GlottHMM can also be used as a general speech analysis tool. In addition to the parameters described above, two additional voice source quantities are extracted: the difference between the first and the second harmonic (H1–H2) [19], describing the spectral tilt of the voice source, and the normalized amplitude quotient (NAQ) [20], describing phonation type. The parameters extracted with GlottHMM are shown in Table 2.

4. Experiments

4.1. Speech material

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

Table 2: Speech features and the number of parameters.

Feature	Number of parameters
Vocal tract spectrum	30
Voice source spectrum	10
Harmonic-to-noise ratio (HNR)	5
Energy	1
Fundamental frequency (F0)	1
H1–H2	1 (only for analysis)
NAQ	1 (only for analysis)

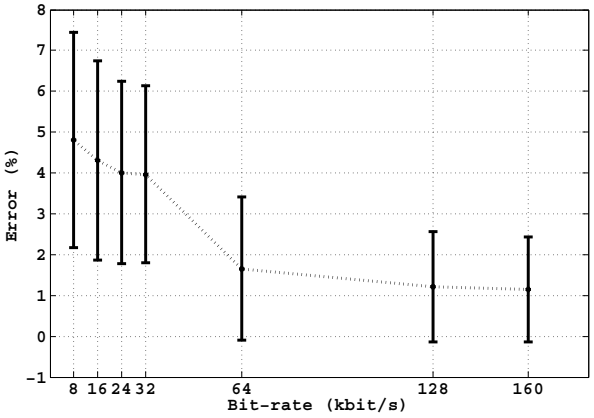

Figure 1: *Relative error of F0 as a function of bit-rate. Data is represented as means and 95% confidence intervals over the two voices (male MV and female HK).*

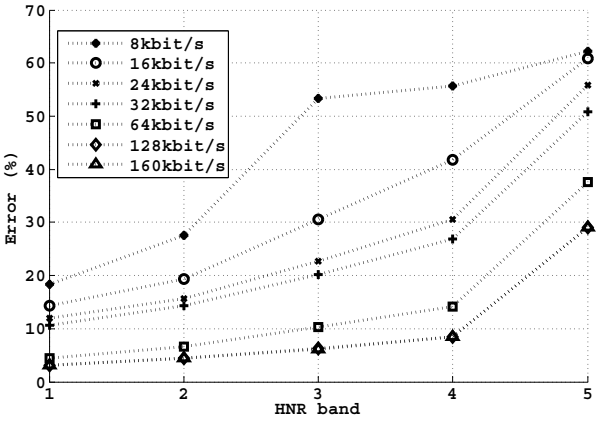

Figure 2: *Relative error of HNR as a function of bit-rate.*

4.2. Objective evaluations of vocoder parameters

The effects of compression (bit-rates shown in Table 1) was evaluated by comparing the vocoder parameters extracted from the MP3-processed sounds to those obtained from the corresponding uncompressed sentences. For each compression rate, the relative error was determined between the parameter value computed from the uncompressed and compressed sound for both speakers. The following five parameters were analyzed: 1) F0, 2) HNR, 3) LSF of the voice source, 4) LSF of the vocal tract, 5) H1–H2, and 6) NAQ.

Figure 1 shows the relative error of F0 as a function of bit-rate. It can observed that the error of F0 for high bit-rates (64 kbit/s or more) is almost negligible and even for low bit-rates the average error is less than 5%. Figure 2 shows the relative error of HNR of each five bands. The error in the high frequency bands is larger than in lower bands, which indicates that the strong harmonic structure of the low-frequency bands is fairly well preserved while the details in the high-frequency bands suffer from compression. Figures 3 and Figure 4 show the relative error of the LSFs of the voice source and the vocal tract, respectively. The figures show that the error is great-

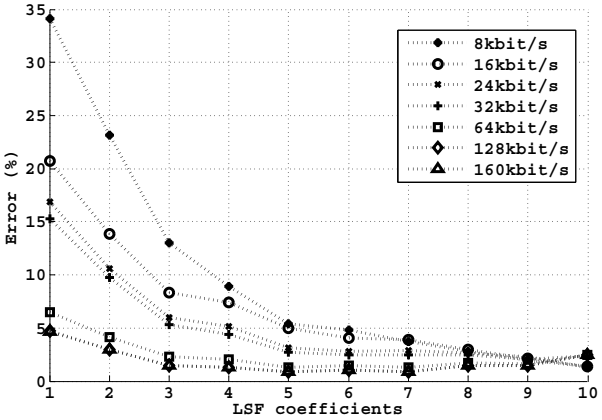

Figure 3: *Relative error of voice source LSFs as a function of bit-rate.*

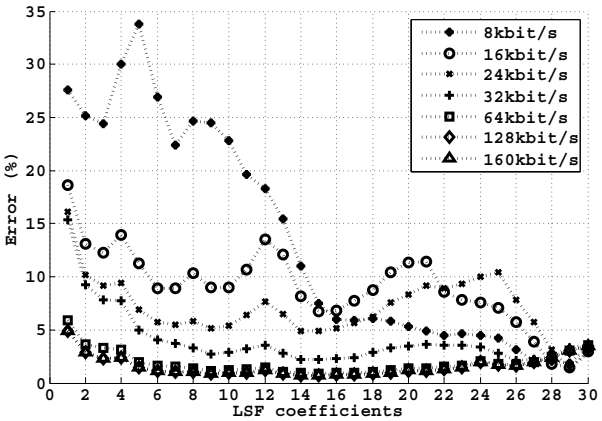

Figure 4: *Relative error of vocal tract LSFs as a function of bit-rate.*

est in low-frequencies. Table 3 shows the correlation coefficient of H1–H2 and NAQ between the original and compressed speech signals. The table shows that correlation declines gradually with decreasing bit-rates. Especially H1–H2 suffers greatly from high compression, while NAQ remains relatively stable.

The objective evaluations of all parameters show a significant increase in error between the bit-rates of 64 and 32 kbit/s. In conclusion, if the bit-rate is 32 kbit/s or less, the compressed acoustic signal is significantly different from the original speech signal, which has a clear effect on the vocoder.

4.3. Evaluation of analysis-synthesis quality

A subjective evaluation was performed to study how the quality of vocoded signals is affected by compression. As subjective evaluations are more laborious than objective ones, only four bit-rates were included in the subjective tests: 128 kbit/s, 32 kbit/s, 16 kbit/s, and 8 kbit/s. From each bit-rate, two sets of signals were selected: 1) compressed signals and 2) vocoded signals. From each category, six randomly selected sentences (3 male and 3 female) were used. A total of 10 native Finnish listeners participated in the evaluations. The sentences were presented in random order to subjects who rated the naturalness of signals on the mean opinion score (MOS) scale, ranging from 1 to 5 (1–completely unnatural, 5–completely natural).

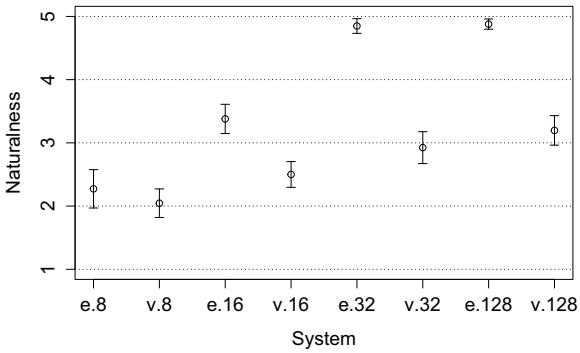

Figure 5: *MOS scores for naturalness where e.8, e.16, e.32, and e.128 refer to encoded speech with bit-rates of 8, 16, 32, and 128 kbit/s, respectively. Vocoded speech is denoted by v.8, v.32, and v.128 corresponding, respectively, to bit-rates of 8, 32, and 128 kbit/s.*

Figure 5 shows the means and 95% confidence intervals of the naturalness ratings. The results show that the subjects rated the 128 kbit/s and 32 kbit/s compressed signals as completely natural, whereas the speech sounds compressed with 16 kbit/s or lower show a clear drop in naturalness. The result suggests that speech, sampled with 16 kHz, can be compressed with a bit-rate as low as 32 kbit/s with very little or without any degradation in quality. A similar trend can also be observed for vocoded signals: vocoded speech corresponding to the bit-rates of 128 kbit/s and 32 kbit/s are rated equal in naturalness, although significantly lower than non-vocoded signals. An interesting observation is that the gap between the compressed signals and vocoded signals is reduced along with the decreased bit-rate.

4.4. Evaluation of HMM-synthesis quality

HMM-based synthetic voices were built with the GlottHMM vocoder. Standard HTS procedure [9, 10] was used for training the voices with the modifications needed to accommodate the increased number of parameters of the vocoder [6]. Compressed speech signals with different bit-rates were used for building 7 voices: full-rate PCM, and 160, 128, 64, 32, 24, 16, and 8 kbit/s. Figure 6 show the average spectra of original speech and HMM-based synthetic voices for the male MV and female HK speakers. Voices with bit-rates of 24 kbit/s or less show a large decrease in magnitude from 5 kHz to 8 kHz and also some deviation from 2 kHz to 5 kHz for both natural compressed and HMM-based voices. However, the low-bit-rate HMM-based voices also show more distortion in their average spectra compared to compressed and original speech.

Table 3: *Correlation coefficient (ρ) of H1–H2 and NAQ between compressed and original uncompressed speech.*

Bit-rate (kbit/s)	**ρ (H1–H2)**	**ρ (NAQ)**
160	0.98	0.94
128	0.97	0.94
64	0.96	0.92
32	0.90	0.88
24	0.85	0.86
16	0.71	0.83
8	0.54	0.81

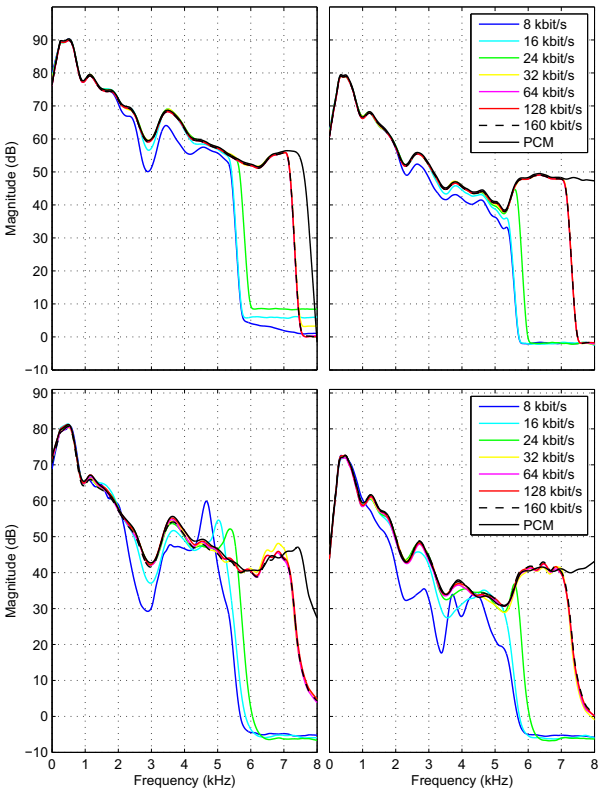

Figure 6: *Long-term average spectra of compressed natural (upper graphs) and synthetic speech (lower graphs) with different bit-rates for the male (left) and female (right) speaker.*

Subjective evaluations were conducted to assess the performance of the HMM-based voices. Four voices were selected for the tests: 1) 128 kbit/s voice, which is considered comparable to training with uncompressed speech since it was rated totally natural in subjective listening tests, and 2) 32 kbit/s, 3) 16 kbit/s, and 4) 8 kbit/s voices, which are heavily compressed. For the subjective evaluations, the same setup as described in Section 4.3 was used. A total of 10 Finnish listeners participated in the MOS naturalness evaluation.

Figure 7 shows the means and 95% confidence intervals of the MOS naturalness ratings. The difference between the naturalness of 128 kbit/s and 32 kbit/s voices is not significant, thus suggesting that the degradations of compression in the 32 kbit/s signal do not affect the training of a HMM-based synthetic voice. The difference between 32 kbit/s and 16 kbit/s voices is greater, indicating degraded quality. The naturalness of 8 kbit/s voice is rated very low, indicating that such high compression rates are not suitable for building HMM-based synthetic voices.

5. Discussion

The results of the experiments show that the vocoder parameters corresponding to high bit-rates such as 160, 128 and 64 kbit/s maintain high fidelity in comparison to the parameters of original, uncompressed speech, whereas the parameters corresponding to low bit-rates (32 kbit/s or lower) are distorted. However, the subjective evaluations indicate that the naturalness of the speech signals is affected by the compression scheme when the

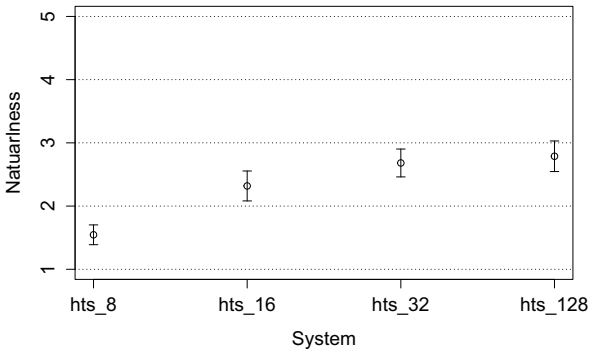

Figure 7: *Naturalness ratings for HMM-based synthetic voices trained with the following bit-rates: 8, 16, 32, and 128 kbit/s.*

bit-rate is 16 kbit/s or less.

In this study, compression of speech signals was done with the MPEG-1 Audio Layer-3 technique, which utilizes a psychoacoustic model for determining masked signals that are less relevant for perception, and which can thus be removed. The error introduced by this compression on F0 is very low, which is irrelevant from a practical point of view. The compression technique, removing less relevant high-frequency content, may explain the large error in high frequency bands of the HNR values. The large error with low-frequency LSFs is partly explained by the higher sensitivity to relative error of the lower coefficients, but the high error of H1–H2 also indicates that low-frequencies are also distorted with high compression rates.

The results of the subjective evaluation of synthetic speech suggest that the effect of compression is smaller when the MP3-processed data are used in training of HMM-based voices. This is confirmed with the bit-rate of 16 kbit/s, where the degradation caused by compression is clearly audible, but only a slight degradation can be observed with the HMM-based voice built from the same signals. This may be due to the statistical training, which averages out occasional audible artefacts but preserves the main characteristics of speech.

6. Conclusions

In this paper, the effects of using MP3-compressed degraded speech in vocoding and HMM-based speech synthesis was studied. Speech signals were encoded with various compression rates and experiments were performed using the GlottHMM vocoder. Both objective and subjective evaluations were used to study the effect of compression on vocoder and HMM-based speech synthesis. Objective evaluation results showed that the vocoder parameters were degraded from encoding with bit-rates of 32 kbit/s or less, which was also confirmed by the subjective evaluation of the vocoder analysis-synthesis quality. Experiments with HMM-based speech synthesis showed that the subjective quality of a synthetic voice trained with 32 kbit/s speech was comparable to a voice trained with uncompressed speech, but lower bit rates induced clear degradation in quality.

7. Acknowledgments

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 and from the Academy of Finland (256961, 135003).

8. References

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

[2] Gonzalez, J. and Cervera, T., "The Effect of MPEG Audio Compression on a Multi-dimensional Set of Voice Parameters", Log. Phon. Vocol., 26(3):124–138, 2001.

[3] van Son, R.J.J.H., "A Study of Pitch, Formant, and Spectral Estimation Errors Introduced by Three Lossy Speech Compression Algorithms", Acta Acustica United With Acustica, 91(4):771–778, 2005.

[4] Zen, H., Tokuda, K. and Black, A.W., "Statistical parametric speech synthesis", Speech Commun., 51(11):1039–1064, 2009.

[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", ISO/IEC 11172-3:1993, International Organization for Standardization, 1993.

[6] Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M. and Alku, P., "HMM-based speech synthesis utilizing glottal inverse filtering", IEEE Trans. on Audio, Speech, and Lang. Proc., 19(1):153–165, 2011.

[7] Tzanetakis, G. and Cook, P., "Sound Analysis Using Mpeg Compressed Audio", Proc. of ICASSP, vol. 2, pp. 761–764, 2000.

[8] LAME encoder, [online] *http://lame.sourceforge.net/*

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

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

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

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

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

[14] Kawahara, H., Masuda-Katsuse, I. and de Cheveigné, A., "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., 27(3–4):187–207, 1999.

[15] Alku, P., "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering", Speech Commun., 11(2–3):109–118, 1992.

[16] Soong, F.K. and Juang, B.-H., "Line spectrum pair (LSP) and speech data compression", Proc. ICASSP, vol. 9, 1984, pp. 37–40.

[17] Marume, M., Zen, H., Nankaku, Y., Tokuda, K. and Kitamura, T., "An investigation of spectral parameters for HMM-based speech synthesis", Proc. Autumn Meeting of Acoust. Soc. of Japan, 2006 (In Japanese).

[18] Moore, B. and Glasberg, B., "A revision of Zwicker's loudness model", ACTA Acustica, 82:335–345, 1996.

[19] Titze, I. and Sundberg, J., "Vocal intensity in speakers and singers", J. Acoust. Soc. Am., 91(5):2936–2946, 1992.

[20] Alku, P., Bäckström, T. and Vilkman, E., "Normalized amplitude quotient for parametrization of the glottal flow", J. Acoust. Soc. Am., 112(2):701–710, 2002.