Efficient quantization of vocoded speech parameters without degradation

Masanori Morise*† and Genta Miyashita‡
* School of Interdisciplinary Mathematical Sciences, Meiji University, Japan
E-mail: mmorise@meiji.ac.jp Tel/Fax: +81-3-5343-8332
† JST, PRESTO, Japan
‡ Integrated Graduate School of Medicine, Engineering, and Agricultural Sciences,
University of Yamanashi, Japan

***Abstract*—In a statistical parametric speech synthesis (SPSS) system with a vocoder, the dimensions of speech parameters need to be reduced, and many SPSS systems have used companded speech parameters. This paper introduces quantization algorithms for 3 speech parameters: fundamental frequency (f_o), spectral envelope, and aperiodicity. In full-band speech (speech with a sampling frequency above 40 kHz), the dimensions of the spectral envelope and the aperiodicity can be reduced to 50 and 5 dimensions based on previous studies. This paper compares the quantization coding without degradation with speech synthesized by the speech parameters without coding. Efficient quantization would be effective for a study that uses graphics processing unit (GPU) computing because recent GPUs support 16-bit floating-point computing. We did two subjective evaluations. The first evaluation determined the appropriate quantization bits in each speech parameter. We obtained the 9 bit values in f_o, 13 bit values in the spectral envelope, and 3 bit values in the aperiodicity. The second evaluation verified the effectiveness of our proposed coding. Since a multiple of eight is generally used for data chunks, we employed the 16 quantization bits for f_o, 16 for the spectral envelope, and 8 for aperiodicity in the evaluation. The results showed that our proposed algorithm achieved almost all the same sound quality as the speech parameters without coding.**

I. Introduction

A vocoder [1] decomposes speech waveforms into the speech parameters related to pitch and timbre. Fundamental frequency (f_o), spectral envelope, and aperiodicity are speech parameters obtained by recent high-quality vocoders [2], [3], [4]. The sound quality of the synthesized speech is almost similar to input speech. Since the speech parameters are related to perceived information, the vocoder is useful for voice conversion techniques such as voice morphing [5] and text-to-speech synthesis [6].

A high-quality vocoder outputs the speech parameters, but the data size is larger than that of the original waveform. In cases where the speech parameters are used for the statistical parametric speech synthesis (SPSS) [7], the memory usage should be reduced for efficient computation. Recent SPSS systems have used the deep neural network (DNN) [8], and a DNN-based singing synthesizer has been proposed [9]. Reducing the bit rate is important because a large amount of speech parameters is required to synthesize natural speech and singing. A new SPSS system such as WaveNet [10] requires no vocoder to generate the waveform, but many SPSS systems still use a high-quality vocoder.

In an SPSS system with a vocoder, the dimensions of each parameter need to be reduced, and many SPSS systems have used companded speech parameters. The recent graphic processing unit (GPU) supports 16-bit floating-point computing and effectively reduces the quantization bits of each parameter to 16 bits. We focus on the quantization bits in all speech parameters. The current quantization bit is fixed at 64 bits. If the quantization exceeds the human auditory system, we can reduce the bit rate by the quantization coding.

We propose coding algorithms for the quantization of each speech parameter without degrading the sound quality. We carried out two subjective evaluations to verify our coding algorithm. The first evaluation determined the appropriate quantization bit of each speech parameter. The second evaluation confirmed that there was only a small difference in sound quality between the speech synthesized by the parameters without coding and the speech parameters with coding.

The rest of this paper is organized as follows. In Section 2, we briefly discuss related works. In Section 3, we explain quantization coding. In Section 4, we detail our two subjective evaluations and their results. In Section 5, we discuss the results and effectiveness of our coding. In Section 6, we conclude with a brief summary and mention our future work.

II. Related works and the concept of proposed coding

Since coding efficiency depends on the sampling frequency, we focus on the coding for the full-band speech, which is speech with a sampling frequency above 40 kHz. The purpose of the study is to achieve speech parameter coding without degrading sound quality. Restricting coding in relation to sound quality is necessary to guarantee the high performance of many applications that use a vocoder.

There are 3 factors related to the bit rate: the frame shift, the number of dimensions, and the quantization bit. A 5-millisecond frame shift has been widely used for many applications. A subjective test [11] has shown that the value was appropriate. Regarding the number of dimensions, the f_o consists of one value per frame, but the other two parameters consist of 1,025 dimensions. These values have been used in the high-quality vocoders such as STRAIGHT [2] and WORLD [4]. Conventional study [12] has shown that the

TABLE I
LOWER AND UPPER LIMITS OF EACH SPEECH PARAMETER.

Parameter	lower limit	upper limit
f_o	1 Hz	4,096 Hz
Spectral envelope	-88	89
Aperiodicity	-60 dB	0 dB

number of dimensions can be reduced to 50 by using the mel-cepstrum [13]. The reduction has been examined [14], and the results showed that the number of dimensions of can be reduced down to 5 in the aperiodicity.

A log domain pulse model [15] and the GlottDNN [16] were proposed as high-quality vocoders with different speech parameters. For example, the GlottDNN uses 111 dimensions per frame, which is larger than high-quality vocoders with coding. The WaveNet vocoder [17] can synthesize the speech from the 3 speech parameters. On the other hand, since the evaluation was carried out using narrow-band speech (16 kHz sampling), the adequacy is difficult to determine.

As just described, many works for efficiently representing the speech parameters in the number of dimensions per frame have been written. μ-law[1] and A-law algorithms were used in the quantization coding for the waveform. They focused on the human auditory system in the waveform amplitude, and making it useful for the waveform quantization. We attempt to efficiently quantize the speech parameters estimated by the high-quality vocoder. The target efficiency is the parameter representations below 16 bit because recent GPUs support 16-bit floating-point computing. Since it is difficult to objectively measure the relationship between the degradation and the speech parameters, we subjectively evaluate the effectiveness.

III. ALGORITHM FOR QUANTIZATION OF THREE SPEECH PARAMETERS

This section explains how to quantize the speech parameters. Table I shows the dynamic ranges of each parameter. We show the argument on why these values were determined.

A. Fundamental frequency quantization

The dynamic range in the f_o was determined from 1 to 4,096 Hz. In the lower limit, the f_o of human speech does not generally indicate 1 Hz. Since the vocoder can work even if such a value is included, the lower limit was set for flexible use. In the higher limit, f_o around 2,000 Hz is often observed in soprano singing and shouted speech. The vocoder can also work for musical instruments, and the f_o of several instruments exceeds that of human speech. In cases where the usage of a vocoder is not limited by human speech, this dynamic range would be reasonable.

The pitch perception of the human auditory system is on the logarithmic frequency in approximately an equidistant manner. Therefore, the quantization was carried out on the logarithmic frequency axis.

B. Spectral envelope quantization

The default quantization bit of the speech parameters is set to 64 bit (double precision floating-point number) in high-quality vocoders. The minimum and maximum values of the 64-bit floating-type number are 2.225074×10^{-308} and 1.797693×10^{308}, respectively. The amplitude of the waveform is limited from -1.0 to 1.0 in the audio file that uses linear pulse code modulation (linear PCM). In this case, the dynamic range in the spectral envelope can be limited to a more narrow range.

The dynamic range of the spectral envelope was based on the single precision floating-point number (32 bit). The minimum and maximum values are 1.175494×10^{-38} and 3.402823×10^{38}, respectively. Their logarithmic values are -87.3365 and 88.7228. Based on the range, we used the dynamic range from -88 to 89. The quantization was carried out on this domain.

This dynamic range can completely cover the dynamic range of a 32-bit floating-type number. Since the spectral envelope is given as the power spectral representation, we can calculate the dynamic range on the dB domain. This range is obtained as -379.3 to 385.3 dB, wide enough to analyze real speech and musical instruments.

C. Aperiodicity quantization

We evaluated using the speech parameters estimated by the WORLD vocoder. Dynamic range of the aperiodicity estimated by WORLD is fixed from -60 to 0 dB. In the synthesis, the value that exceeds this range is rounded up/down to fix the range. Therefore, the dynamic range used for the quantization was set to the same range according to the specification of WORLD. The quantization was carried out in an equidistant manner on the dB axis.

IV. SUBJECTIVE EVALUATIONS

The evaluations consisted of two subjective listening tests. The evaluations showed that the sound quality of the speech synthesized by coded speech parameters is almost the same as that of the speech synthesized by the original speech parameters. The adequacy and the efficiency of our coding are discussed based on results of the evaluations.

A. Vocoder used in the evaluations

In both evaluations, we used WORLD (D4C edition [14]) as a high-quality vocoder. Since it has several estimators for each parameter, we used Harvest [18] to estimate the f_o, CheapTrick [19], [20] to estimate the spectral envelope, and D4C [14] to estimate the aperiodicity. The frame shift was set to 5 ms, and other parameters were set to their defaults.

Before evaluating, we aurally checked the sound quality of the synthesized speech and confirmed there were no fatal errors that could degrade it. Doing so enabled us to purely measure the degradation with the quantization coding. The number of dimensions in the spectral envelope was 1,025, but we used mel-cepstrum of 50 dimensions based on the previous study [12].

[1] https://www.cisco.com/c/en/us/support/docs/voice/h323/8123-waveform-coding.html

TABLE II
EXPERIMENTAL CONDITIONS IN THE FIRST EVALUATION.

Evaluation protocol	
Method	MUSHRA-based evaluation
Number of subjects	10 people
Environment and equipment	
Environment	Soundproof room
Background noise	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 men and 2 women)
Number of stimuli	20 (5 words per speaker)
Kind of speech	4-mora words including consonants
Sampling	48 kHz/16 bit
Quantization bits used in each speech parameter	
f_o	7, 8, 9, and 10 bit
Spectral envelope	11, 12, 13, and 14 bit
Aperiodicity	2, 3, 4, and 5 bit

B. First evaluation to determine the appropriate quantization bits

1) Experimental conditions: Table II represents the conditions in the first evaluation. The evaluation was based on multiple stimuli with hidden reference and anchor (MUSHRA) defined by ITU-R recommendation BS.1534-3. In the evaluation, the subjects scored the speech stimuli on a scale of 0 to 100 (full marks) by using a graphical user interface (GUI) since the MUSHRA-based evaluation can generally evaluate smaller differences than the mean opinion score (MOS). We used a soundproofed room with an A-weighted SPL of 18 dB, and 10 people with normal hearing abilities participated in the evaluation. We used a set of headphones (SENNHEISER HD650) for the evaluation.

We conducted 3 listening tests to determine the appropriate quantization bits for each speech parameter. The speech stimuli used for the evaluations were 20 words spoken by 2 men and 2 women. The speech consisted of Japanese 4-mora words that included consonants. The sampling frequency was 48 kHz and the quantization bit was 16 bit.

In each test, the participants evaluated 5 speech stimuli at the same time by using the GUI. The reference was resynthesized speech with the original speech parameters (64 bit). Four other conditions were determined by exploratory listening tests. The speech stimuli were randomized and reproduced to the subjects through the headphones.

2) Results and obtained quantization bits: Figs. 1, 2, and 3 show the experimental results. In all the figures, the vertical axis represents the average scores under each condition. The error bar represents the 95% confidence interval. Results of the statistical analysis indicated that there was no significant difference between the original (64 bit) and the highest quantization bits in each speech parameter. Since a multiple of eight is generally used for data chunks, we employed the 16 quantization bits for f_o, 16 for the spectral envelope, and 8 for aperiodicity.

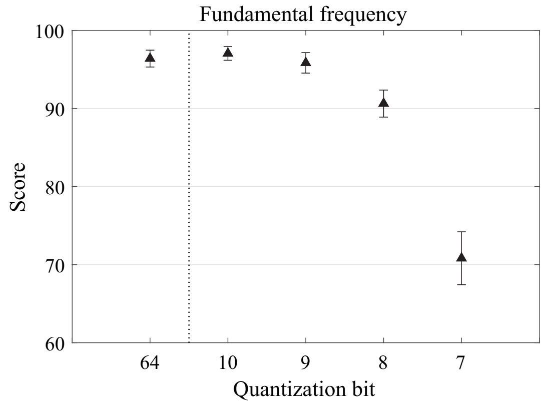

Fig. 1. Relationship between the sound quality and the quantization bits of f_o.

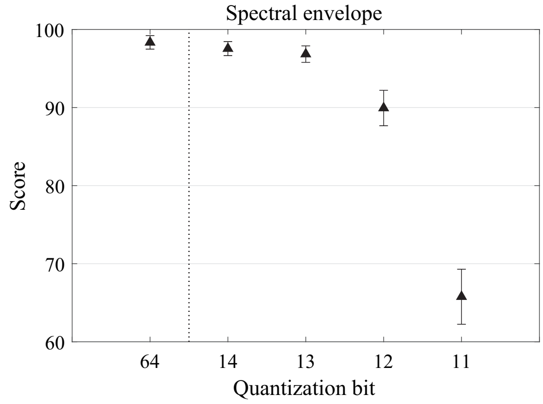

Fig. 2. Relationship between the sound quality and the quantization bits of spectral envelope.

C. Second evaluation to compare the sound quality

1) Experimental conditions: In the second evaluation, we employed a Thurstone's paired comparison test instead of the MUSHRA-based evaluation. MUSHRA-based evaluation requires the reference as the best sound quality, and the sound quality of coded speech must be lower than that of reference. Speech synthesized with coded speech parameters has often been preferred over that with original speech parameters [12]. The Thurstone's paired comparison test is a reasonable way to evaluate sound quality compared with the MUSHRA-based evaluation.

Table III represents the conditions in the second evaluation. In the experiment, we used 3 conditions. The first condition was the speech parameters without coding. Since the speech parameters were estimated by the WORLD vocoder, the numbers of dimensions were 1 in the f_o, 1,025 in the spectral envelope, and 1,025 in the aperiodicity. The quantization bit was 64 bits in all parameters. The second condition was the speech parameters with coding based on previous studies [12], [14]. The number of dimensions were 1 in the f_o with 16 quantization bits, 50 in the spectral envelope with 16 quantization bits, and 5 in the aperiodicity with 8 quantization

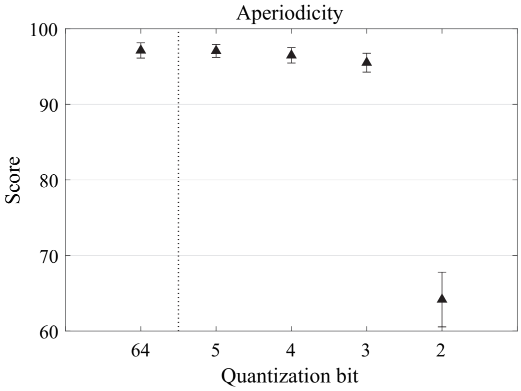

Fig. 3. Relationship between the sound quality and the quantization bits of aperiodicity.

TABLE III
EXPERIMENTAL CONDITIONS IN THE SECOND EVALUATION.

Conditions	
Method	Thurstone's Paired Comparison
Number of subjects	20 people
Number of stimuli	40 (10 words per speaker)
Quantization bits and dimensions per frame	
f_o	16 bit (1 dimension)
Spectral envelope	16 bit (50 dimensions)
Aperiodicity	8 bit (5 dimensions)

bits.

This evaluation expanded on the first evaluation to improve the reliability of the results in the first evaluation. We increased the number of stimuli from 20 to 40 words and the number of subjects from 10 to 20 people. In the evaluation, the subjects listened to the two speech stimuli and then told us which one they preferred. Two stimuli consisted of the same words but different, randomly selected conditions. Therefore, each subject evaluated the 240 pairs (40 speech stimuli × 6 combinations) in the evaluation.

2) Result: Fig. 4 shows the experimental result. The horizontal axis represents the percentage related to the sound quality. The result showed that our coding achieved almost all the same quality compared with the condition without coding (w/o coding: 48% and our coding: 52%). The difference between the baseline and our coding was also low (Baseline: 52.1% and our coding: 47.9%).

We calculated the subjective scales related to the sound quality from the values in Fig. 4. Subjective scales in three conditions were -0.06 (w/o coding), 0.00 (proposed coding), and 0.06 (baseline). These results showed that the coding has little influence on the sound quality.

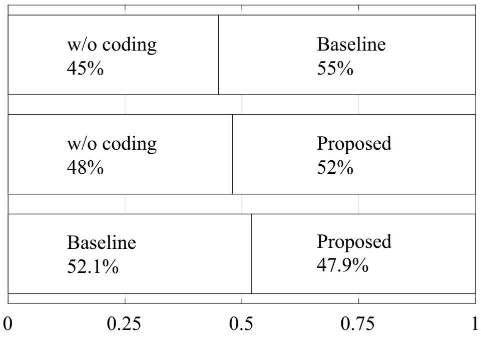

Fig. 4. Experimental results of the Thurstone's paired comparison test.

V. DISCUSSION

First, we explain the coding efficiency compared with the waveform and original speech parameters. Then, we discuss the effectiveness of our coding and future work to achieve representation that is more efficient.

A. Coding efficiency by the proposed coding

Since the speech used in the evaluation was 48 kHz/16 bit, its bit rate was 768 kbps. The bit rate of w/o coding was 26.25 Mbps (200 frame/s × (1 + 1,025 + 1,025) dim × 64 bit). The bit rate of the baseline was 716.8 kbps (200 frame/s × (1 + 50 + 5) dim × 64 bit). The baseline was a slightly low bit rate compared with the original waveform.

The bit rate of our coding was 171.2 kbps (200 frame/s × ((1 + 50) dim × 16 bit + 5 dim × 8 bit)). The coding efficiency was 23.8% without degradation compared with the baseline. The frame shift was fixed to 5 ms and not considered in the evaluation, but the bit rate was reduced compared with the original waveform.

B. Effectiveness of the proposed coding

Since an SPSS system that uses the vocoder uses the speech parameters, this coding would be effective for the SPSS study. Our coding achieved the 16-bit representation in all speech parameters, which means that the GPU computing with 16-bit would be possible by using our coding. Our coding would cut the computational cost and memory usage.

C. Future work

The dynamic ranges of each speech parameter were not optimized because they were set to avoid the values of speech parameters that do not exceed the range. Our coding in f_o was processed on the logarithmic frequency axis. Another frequency scale such as a mel scale could be more effective for achieving the coding efficiency. Bit distribution on the non-linear axis is an important task that requires study to show the relationship between the human auditory system and the speech parameters.

Limiting the dynamic range would achieve better coding efficiency. In the f_o, we can use a preprocessing of speech parameters to limit the dynamic range of the input as usage.

In the spectral envelope, since the dynamic range of the amplitude of linear PCM is limited from -1.0 to 1.0, the maximum value would be estimated theoretically. These adjustments can achieve more efficient coding.

VI. Conclusion

We proposed quantization coding algorithms for the vocoded speech parameters. We carried out two subjective evaluations to show the appropriate quantization bits for each speech parameter. Therefore, we could achieve similar sound quality compared with speech without coding by using 16 quantization bits for f_{o}, 16 for the spectral envelope, and 8 for the aperiodicity. Since the default quantization bit was 64 bits, the compression ratio of the speech parameters achieved 23.8%.

The next step of our study is to improve the coding efficiency. our algorithm covered the excess range in the f_{o} (from 1 to 4,096 Hz) and spectral envelope (from -379.3 to 385.3 dB). Determining the appropriate frame shift is also important. Many researchers have used 5 ms, but whether this value is the best in full-band speech is unclear. Efficient representations for the speech parameters without degradation is our important task.

VII. Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP16H01734 and JP16H05899, and JST PRESTO Grant Number JPMJPR18J8, Japan.

References

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

[2] 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.

[3] 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.

[4] 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.

[5] H. Kawahara, R. Nisimura, T. Irino, M. Morise, T. Takahashi, and H. Banno, “Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown,” *in Proc. of ICASSP2009*, pp. 3905–3908, 2009.

[6] A. Kain and M. W. Macon, “Spectral voice conversion for text-to-speech synthesis,” *in Proc. ICASSP 1998*, pp. 285–288, 1998.

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

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

[9] M. Blaauw and J. Bonada, “A neural parametric singing synthesizer modeling timbre and expression from natural songs,” *Applied Science*, vol. 7, no. 12, pp. 23–page, 2018.

[10] 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.

[11] T. Kitamura, S. Imai, C. Furuichi, and T. Kobayashi, “Speech analysis-synthesis system and quality of synthesized speech using mel-ceptrum,” *IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences*, vol. J68-A, no. 9, pp. 957–964, 1985 (in Japanese).

[12] M. Morise and G. Miyashita, “Low-dimensional representation of spectral envelope without deterioration for full-band speech analysis/synthesis system,” *in Proc. INTERSPEECH 2017*, pp. 409–413, 2017.

[13] 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.

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

[15] G. Degottex, P. Lanchantin, and M. Gales, “A log domain pulse model for parametric speech synthesis,” *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 26, no. 1, pp. 57–70, 2018.

[16] 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.

[17] K. Kobayashi, T. Hayashi, A. Tamamori, and T. Toda, “Statistical voice conversion with WaveNet-based waveform generation,” *in Proc. INTERSPEECH 2017*, pp. 1138–1142, 2017.

[18] M. Morise, “Harvest: A high-performance fundamental frequency estimator from speech signals,” *in Proc. INTERSPEECH2017*, pp. 2321–2325, 2017.

[19] ——, “CheapTrick, a spectral envelope estimator for high-quality speech synthesis,” *Speech Communication*, vol. 67, pp. 1–7, 2015.

[20] ——, “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.