RANGE-CONSTRAINED PHASE RECONSTRUCTION FOR RECOVERING TIME-DOMAIN SIGNAL FROM QUANTIZED AMPLITUDE AND PHASE SPECTROGRAM

Sho Sato,
Graduate School of
Information Science and Technology
The University of Tokyo
Tokyo, Japan
sato@hil.t.u-tokyo.ac.jp

Nobutaka Ono,
National Institute of Informatics
Tokyo, Japan
onono@nii.ac.jp

Yutaka Kamamoto,
NTT Communication Science Laboratories
Nippon Telegraph and Telephone Corporation
Kanagawa, Japan
kamamoto.yutaka@lab.ntt.co.jp

Shigeki Sagayama,
Graduate School of
Information Science and Technology
The University of Tokyo
Tokyo, Japan
sagayama@hil.t.u-tokyo.ac.jp

ABSTRACT

This paper describes a novel algorithm for recovering time-domain signal from quantized amplitude and phase spectrogram, which is applicable for spectrogram-based audio coding. In order to obtain a better quality sound, a phase reconstruction technique is first applied with constraint for keeping phase in each time-frequency bin within each quantization range, and then, time-domain signal is recovered by the standard inverse short-time Fourier transform. Experimental evaluation based on the objective PEAQ measure shows that the proposed range-constrained phase reconstruction is effective for improving the sound quality.

1. INTRODUCTION

Portable digital audio players enable us to enjoy listening to music anywhere. To reduce the size of files, the music data is usually encoded with lossy audio coding, typically the MPEG-1/2 Audio layer 3 (MP3) or MPEG-2/4 Advanced Audio Coding (AAC) [1, 2] standards. These state-of-the-art codecs exploit the redundancies of audio signals, psycho-acoustic models, the statistics of the parameters, and so on. These coding schemes have been well investigated and are now highly sophisticated, and thus, the performance has gradually reached saturation.

On the other hand, many audio signal processing methods are based on spectrogram domain. Generally, the spectrogram can be obtained by the Short-Time Fourier Transform (STFT). Unlike the Discrete Cosine Transform (DCT) or the Modified Discrete Cosine Transform (MDCT), STFT coefficients consist of amplitude and phase, which represent power and timing information separately, and the amplitude spectrogram represents the music signal structure (such as repetition) well. Furthermore, in the music signal processing field, Non-negative Matrix Factorization (NMF) [3] has recently been investigated as a powerful tool to efficiently decompose an amplitude spectrogram into a finite number of spectral bases and corresponding temporal activations. It is convenient for non-real-time audio coding applications and enables us to exploit long-term redundancies of music signals. Also, the spectrogram representation is suitable for psycho-acoustic models, and thus, the minimization of perceptual distortion in NMF representation has previously been reported in the literature [5].

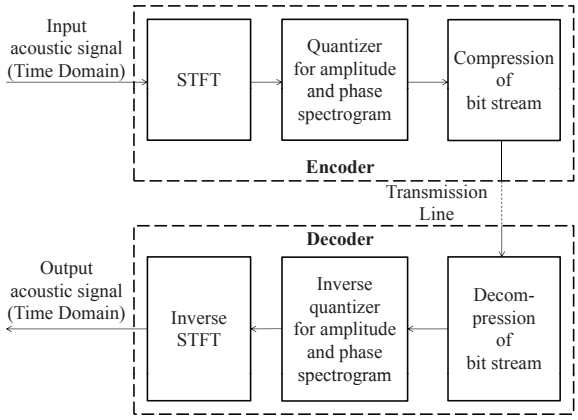

Figure 1: Flowchart of the spectrogram-based audio coding scheme

Motivating from significant advances in music signal processing with modern spectrogram-based techniques such as NMF, we have also explored a spectrogram (STFT)-based audio coding scheme, an overview of which is shown in Fig. 1. A pioneering work using this scheme was done by Mahieux, *et al.* [6], and it was also recently investigated for the object-based audio coding [7]. Generally in the spectrogram-based audio coding, the amplitude and the phase of STFT are quantized in the encoder. The contribution of this paper is to examine how to recover the time-domain signal from the quantized amplitude and phase in the decoder. Basically, the amplitude and the phase in an STFT representation are not mutually independent due to the redundancy originating

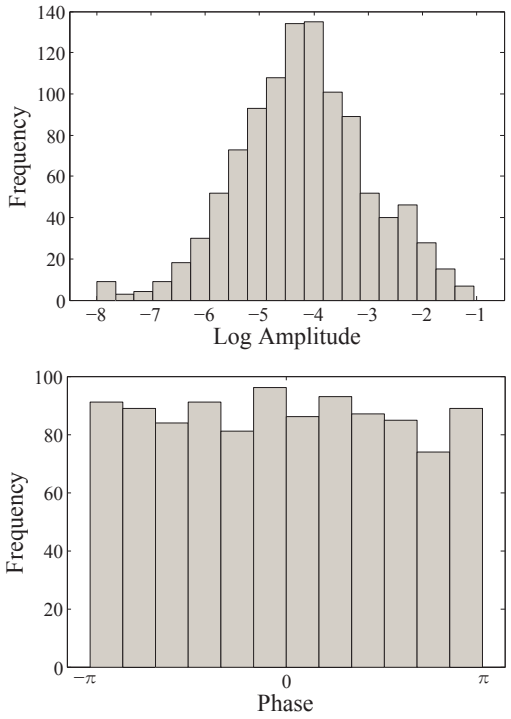

Figure 2: Histograms of logarithmic amplitudes and phases at a frequency band

from the overlapped window analysis of STFT. Phase reconstruction [8, 9, 10] is a technique to reconstruct consistent phases from only the amplitude spectrogram using this redundancy. For obtaining better quality sound, we discuss an attempt to apply this technique in the decoder. Unlike the standard phase reconstruction problem, quantized phase information is available in this context. Therefore, the phase reconstruction algorithm is also modified to keep the estimated phase within the quantization range. An experimental evaluation based on the objective perceptual evaluation of audio quality (PEAQ) measure [13] shows that the proposed range-constrained phase reconstruction improves the sound quality.

2. SPECTROGRAM-BASED AUDIO CODING

2.1. Quantization of amplitude and phase spectrogram

Let Y_{km} be the STFT representation of an input discrete signal x_n, where k and m are the indices of the time frame and the frequency, respectively. It can be expressed as

$$Y_{km} = \exp(A_{km} + j\phi_{km}), \tag{1}$$

where A_{km} and ϕ_{km} represent the logarithmic amplitude and phase of a complex-value Y_{km}. We here define ϕ_{km} from $-\pi$ to π. In order to focus on how to recover the time domain signal from the quantized amplitude and phase in the decoder, we simply apply a uniform quantization to both the logarithmic amplitude and the phase in the encoder in this paper. The introduction of a psycho acoustic model to the encoder is also a significant and necessary task in the STFT-based audio coding too, and we will address it in future work.

Figure 2 shows examples of histograms of logarithmic amplitude A_{km} and phase ϕ_{km} at the m-th frequency bin for a piece of classical music, where the vertical axis represents a number of time-frequency bins within each range. The logarithmic amplitude forms a unimodal distribution and its mean depends on the frequency, whereas the phase forms a uniform distribution at all frequencies. Hence, the quantizing range and step size can vary for the amplitude but are fixed for the phase. The quantization can be represented as follows:

$$\hat{A}_{km} = \left\lceil \frac{A_{km} - \bar{A}_m}{\Delta A_m} + \frac{1}{2} \cdot 2^a \right\rceil, \tag{2}$$

$$\hat{\phi}_{km} = \left\lceil \frac{\phi_{km}}{\Delta\phi} + \frac{1}{2} \cdot 2^p \right\rceil, \tag{3}$$

where $\hat{A}_{km}$ and $\hat{\phi}_{km}$ are the quantized logarithmic amplitude and phase, respectively, and $\lceil \cdot \rceil$ represents the ceiling function. $\bar{A}_{km}$ denotes the mean of A_{km} over all ks, while a and p are the numbers of bits assigned for amplitude and phase, respectively. ΔA_m and $\Delta\phi$ denote the quantization step sizes for the amplitude and the phase, respectively, which are calculated as

$$\Delta A_m = 2 \cdot 3\sigma_m / 2^a, \tag{4}$$

$$\Delta\phi = 2\pi / 2^p, \tag{5}$$

where σ_m denotes the standard deviation of A_{km} around $\bar{A}_{km}$ in terms of k. Consequently, A_{km} in the range $[-3\sigma_m + \bar{A}_m, 3\sigma_m + \bar{A}_m]$ is quantized using a bits. ϕ_{km} in the range $[-\pi, \pi]$ is quantized using p bits which depend on amplitude because higher amplitudes cause larger perceptual distortion when all phases are quantized using the same bits. In future, a more sophisticated quantization technique such as noise shaping [6] or novel polar quantization [11] will be evaluated.

2.2. Recovering time-domain signal from quantized amplitude and phase

The purpose of the decoder in this framework is to recover the time-domain signal from the quantized amplitudes and phases. We aimed to improve the sound quality by applying the phase reconstruction technique [8]. The following three methods are compared in this paper.

2.2.1. Baseline

The simplest way to recover the time-domain signal from the quantized amplitudes and phases is by simply applying the inverse quantization, which replaces the code by the center point of the quantization range. It can be expressed as:

$$A'_{km} = \left(\hat{A}_{km} - \frac{1}{2} \cdot 2^a - \frac{1}{2} \right) \cdot \Delta A_m + \bar{A}_m, \tag{6}$$

$$\phi'_{km} = \left(\hat{\phi}_{km} - \frac{1}{2} \cdot 2^p - \frac{1}{2} \right) \cdot \Delta\phi, \tag{7}$$

where A'_{km} and ϕ'_{km} are the dequantized logarithmic amplitude and phase, respectively. Then, the complex spectrogram can be reconstructed as

$$Y'_{km} = \exp(A'_{km} + j\phi'_{km}), \tag{8}$$

and the time-domain signal is recovered as

$$\boldsymbol{x} = \text{invSTFT}(\boldsymbol{Y'}), \tag{9}$$

where $\boldsymbol{Y'}$ denotes the set of Y'_{km} and $\boldsymbol{x}$ denotes the recovered time-domain signal.

2.2.2. Phase reconstruction

The quantized amplitudes and phases include quantization errors, and thus they may be inconsistent, i.e. if we apply the STFT again to the recovered time-domain signal $\boldsymbol{x}$, the resulting spectrogram may differ from the original (directly-dequantized) spectrogram Y'_{km}. Based on the assumption that the auditory perception is more sensitive to the amplitude than to the phase, better sound quality can be obtained by fixing the amplitudes and updating the phases not to change them through the inverse STFT and STFT, which can be performed by a standard phase reconstruction technique. The procedure is summarized as follows.

With the dequantized values as the initial values ($\boldsymbol{Y} = \boldsymbol{Y'}$), the following update rules are iteratively applied.

$$\boldsymbol{x} = \text{invSTFT}(\boldsymbol{Y}), \quad (10)$$

$$\boldsymbol{Y} = \text{STFT}(\boldsymbol{x}), \quad (11)$$

$$\phi_{km} = \angle Y_{km}. \quad (12)$$

2.2.3. Range-constrained phase reconstruction

When we directly apply the conventional phase reconstruction technique as describe in the previous subsection, the phases move from the initial value (direct dequantized value) and sometimes go out of the quantization range such as (1') or (1") shown in Fig. 3. This is not desirable because the original phase exists in the quantization range before quantization. Therefore, another reasonable way to update the phase is to explicitly exploit the quantization range as a constraint and to draw the phase back to the closest border of the range, as in a clipping operation if the estimated phase goes out of the range, which is shown as (2') or (2") in Fig. 3. The flowchart is also shown in Fig. 4. It can be denoted as the range-constrained phase reconstruction. The procedure is summarized as follows.

With the dequantized values as the initial values ($\boldsymbol{Y} = \boldsymbol{Y'}$), the following update rules are iteratively applied.

$$\boldsymbol{x} = \text{invSTFT}(\boldsymbol{Y}) \quad (13)$$

$$\boldsymbol{Y} = \text{STFT}(\boldsymbol{x}) \quad (14)$$

$$\phi_{km} = \begin{cases} \phi'_{km} + \dfrac{\Delta\phi}{2} & (\angle\left(\dfrac{Y_{km}}{\exp(j\phi'_{km})}\right) > \dfrac{\Delta\phi}{2}) \\ \phi'_{km} - \dfrac{\Delta\phi}{2} & (\angle\left(\dfrac{Y_{km}}{\exp(j\phi'_{km})}\right) < -\dfrac{\Delta\phi}{2}) \\ \angle Y_{km} & (\text{otherwise}) \end{cases} \quad (15)$$

where ϕ'_{km} is the dequantized phase and $(\phi'_{km} - \Delta\phi/2, \phi'_{km} + \Delta\phi/2)$ represents the quantization range of ϕ_{km}.

3. EXPERIMENTAL EVALUATION

3.1. Experimental conditions

In our experiments we have used constant bit rate of 128 kbps with 8 bits used to encode the amplitude and the phase. Table 1 shows the possible modes for allocating the bits. For the testing signals, we have selected 20 excerpts with 20s length, where five pop music and five classical music were selected from RWC

Table 1: Assigned bits for amplitude and phase.

Amplitude [bit]	0	1	2	3	4	5	6	7	8
Phase [bit]	8	7	6	5	4	3	2	1	0

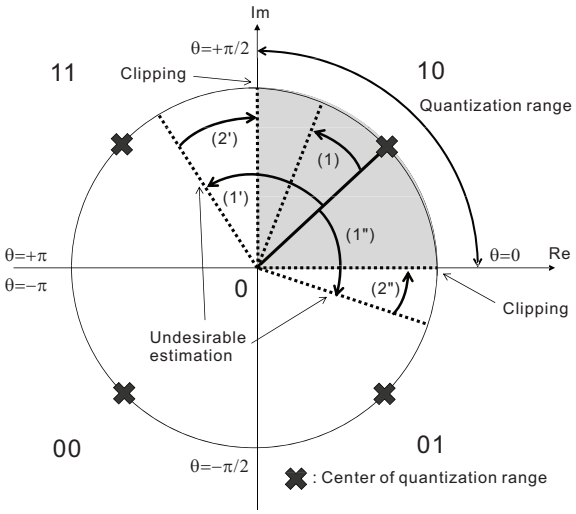

Figure 3: Range-constrained phase reconstruction in the complex plane

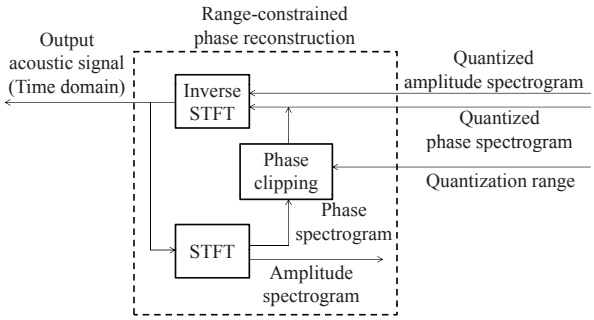

Figure 4: Flowchart of range-constrained phase reconstruction scheme.

database [12], and five male voices and five female voices were selected from ATR Japanese speech database (Set B). All of them were converted to monaural 16 kHz sampling. The frame size was set to 512 samples with a 50% overlap. A Hamming window was used as a windowing function, and was applied before calculating the STFT. The phase reconstruction or the range-constrained phase reconstruction was iteratively performed by 200 times. We used the PEAQ measure to obtain objective, repeatable evaluation results. Its objective difference grade (ODG) scores range from -4 (Very annoying) to 0 (Imperceptible).

3.2. Experimental results

Figures 5 and 6 plot the results of using the three different phase reconstruction methods for music (pop and classical) and voices (male and female), respectively, where the averaged ODG score for 10 excerpts are displayed.

In the baseline method (i.e., direct conversion from dequantized amplitude and phase spectrogram with inverse STFT), the best score for the 8-bit encoding was achieved at (amplitude, phase) = (5, 3) [bits]. We can see that the normal phase reconstruction does not always improve the ODG scores. When the bit assigned to amplitude was more than or equal to 6 bits, the phase reconstruction improved the scores. However, when the bit as-

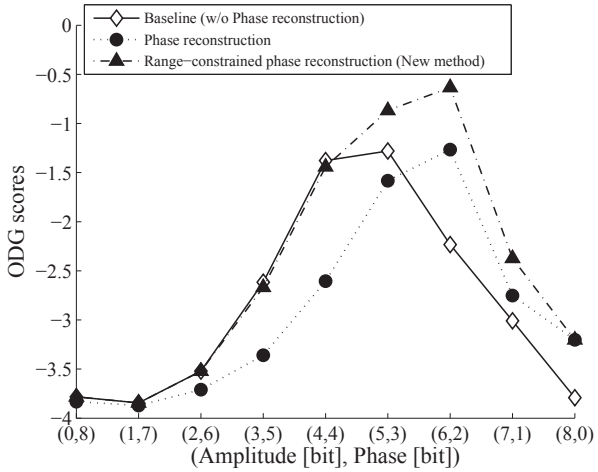

Figure 5: Average ODG scores of PEAQ over ten music excerpts for three different inverse quantization methods

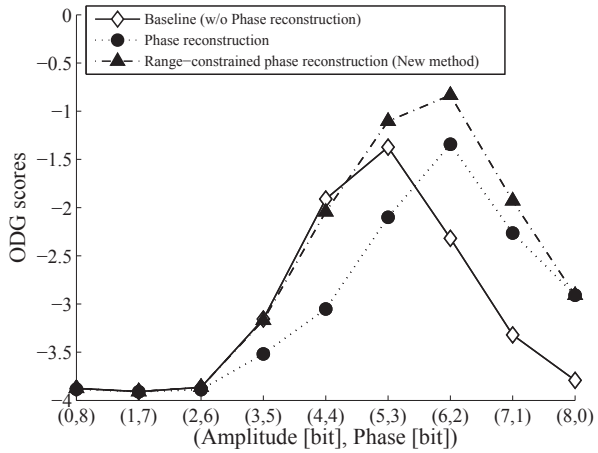

Figure 6: Average ODG scores of PEAQ over ten speech for three different inverse quantization methods

signed to amplitude was less than or equal to 5 bits, the scores decreased. This may be because the amplitude in such bit assignments includes large quantization errors, and the estimated phase from the amplitude may also be unreliable. In contrast, the range-constrained phase reconstruction demonstrated high performance compared with both the baseline and the normal phase reconstruction methods. In this case, the best score was achieved at (amplitude, phase) = (6, 2) [bits].

We compared these results with MP3, one of the standard audio codecs. When MP3 in iTunes10 was used, the resulting averaged ODG scores for the ten music pieces were -0.0181, -0.8913, and -2.3142 for 128 kbps, 64 kbps, and 32 kbps, respectively. Therefore, the proposed range-constrained phase reconstruction at (amplitude, phase) = (6, 2) [bits] assignment (128 kbps) performed comparably to MP3 codec at 64 kbps. However, note that the main focus of this paper is the decoder of STFT-based codec, and the proposed method includes neither the psycho-acoustic model nor the compression algorithm for the amplitude and phase spectrogram yet. The bit rate of the STFT codec will be improved by introducing them in future work.

4. CONCLUSIONS

In this paper, aiming to developing a means of spectrogram-based audio coding, we have presented a new algorithm for recovering time-domain signal from quantized amplitude and phase spectrograms. The experimental results clearly show the proposed range-constrained phase reconstruction was effective in improving sound quality. In future work, we will introduce the psycho-acoustic model into the encoder and investigate an effective compression algorithm for amplitude spectrogram by exploiting NMF or other signal processing techniques to further improve the compression ratio and audio quality.

5. ACKNOWLEDGMENTS

This work was partially supported by JSPS (Japan Society for Promotion of Science) Grant-in-Aid for Challenging Exploratory Research No. 23650083.

6. REFERENCES

[1] M. Bosi and R. E. Goldberg, Introduction to Digital Audio Coding and Standards, Kluwer Academic Publisher, 2003.

[2] A. Spanias, T. Painter, and V. Atti, Audio Signal Processing and Coding, Wiley Interscience, 2007.

[3] D. D. Lee and H. S. Seung, "Algorithms for Non-Negative Matrix Factorization," in *Proc. NIPS*, pp. 556–562, 2000.

[4] A. Griffin, T. Hirvonen, A. Mouchtaris, and P. Tsakalides, "Encoding the Sinusoidal Model of an Audio Signal Using Compressed Sensing," in *Proc. ICME.* , pp. 153–156, 2009.

[5] J. Nikunen and T. Virtanen, "Noise-to-mask Ratio Minimization by Weighted Non-negative Matrix Factorization," in *Proc. ICASSP*, pp. 25–28, 2010.

[6] Y. Mahieux, J. P. Petit, and A. Charbonnier, "Transform Coding of Audio Signals Using Correlation Between Successive Transform Blocks," in *Proc. ICASSP*, pp. 2021-2024, May 1989.

[7] J. Nikunen and T. Virtanen, "Object-Based Audio Coding Using Non-Negative Matrix Factorization for the Spectrogram Representation," in *Proc. 128th Audio Engineering Society Convention*, London, UK, 2010.

[8] D. W. Griffin and J.S. Lim, "Signal Estimation from Modified Short-time Fourier Transform," *IEEE Trans. ASSP.*, vol. 32, no. 2, pp. 236–243, 1984.

[9] J. Le Roux, N. Ono, and S. Sagayama, "Explicit Consistency Constraints for STFT Spectrograms and Their Application to Phase Reconstruction," in *Proc. SAPA*, Sept. 2008.

[10] N. Sturmel and L. Daudet, "Signal Reconstruction from STFT Magnitude: A State of the Art," in *Proc. DAFx*, pp. 375-386, Sept. 2011.

[11] R. Vafin and W. B. Kleijn, "Entropy-constrained Polar Quantization and Its Application to Audio Coding," *IEEE Trans. SAP.*, vol. 13, no. 2, pp. 220–232, Mar. 2005.

[12] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC Music Database: Music Genre Database and Musical Instrument Sound Database," in *Proc. ISMIR*, pp. 229-230, Oct. 2003.

[13] Perceptual Evaluation of Audio Quality (PEAQ), `http://www-mmsp.ece.mcgill.ca/Documents/Software/index.html`