Experimental Evaluation of WaveRNN Predictor for Audio Lossless Coding

Shota Amada[1], Ryosuke Sugiura[2], Yutaka Kamamoto[2], Noboru Harada[2], Takehiro Moriya[2],
Takeshi Yamada[1] and Shoji Makino[1]

[1] University of Tsukuba
1–1–1 Tennoudai, Tsukuba-shi, Ibaraki, 305–0006, Japan
E-mail: s1720651@s.tsukuba.ac.jp

[2] NTT Communication Science Laboratories,
Nippon Telegraph and Telephone Corporation,
3–1 Morinosatowakamiya, Atsugi-shi, Kanagawa,
243–0198, Japan

Abstract

This paper describes a new scheme of speech and audio lossless coding. MPEG-4 Audio Lossless Coding (ALS) is the international standard lossless compression method of audio signals. It uses linear predictor and compresses the signal by converting the signal into information such as prediction residuals and prediction coefficients. In the compressed signal with MPEG-4 ALS, the prediction residuals occupies a large amount of information. Improving the performance of the predictor is directly related to improving compression performance. Using non-linear predictors in lossless coding has the possibility to perform flexible prediction and improve prediction performance. WaveRNN is the deep neural networks performing a non-linear prediction. The outputs of WaveRNN are the predicted probability distribution of the target sample. Arithmetic coding can generate an optimal code for a set of arbitrary symbols and probabilities. WaveRNN is trained to minimize the bitrate after encoding when using arithmetic coding for the output of WaveRNN. This paper proposes the scheme of speech and audio lossless coding that combines WaveRNN with arithmetic coding. Experimental evaluation confirmed that the developed method reduced around 0.7 bits per sample compared to MPEG-4 ALS in speech coding. DNN-based speech and audio coding techniques are expected to transcend the international standard technologies.

1. Introduction

Broadband networks and digital devices have developed and there is an increasing demand for high quality audio signals. The amount of information required for archive and broadcast is dramatically increasing. The compression method with less deterioration of quality is required because the demand of using large amount of information with high quality, such as high-resolusion audio is increasing. (Generally, high-resolution audio is the audio files that have a higher sampling frequency and/or bit depth than that of Compact Disc Digital Audio, which is specified at 16 bit/44.1 kHz.) Lossless compression meets the demand as it can reduce the rate while keeping the perfect reconstruction of the signal. Therefore, it plays an important role for the high quality sound transmission such as high-resolusion audio. Considering to improve the performance of lossless compression, the compression ratio and the calculation amount attracts attention as the performance index because the original signal is losslessly reconstructed. In this paper, we aim to improve the encoding performance about the compression ratio in lossless compression method.

MPEG-4 Audio Lossless Coding (ALS) is the international standard lossless compression method of audio signals[1, 2, 3]. It can handle many types of input signals including high-resolution audio such as one with sampling frequency up to about 4 GHz, quantization bitdepth up to 32 bit and up to 65536 channels. It uses linear prediction analysis and the performance of linear predictor affects its compression ratio. The studies to improve the linear predictor in MPEG-4 ALS has been conducted so far[4]. According to these studies, it seems to exist signals that are difficult for the conventional linear predictor to predict. Using non-linear predictors in lossless coding has the possibility to perform flexible prediction that is not limited to linear prediction and improves prediction performance.

WaveNet and WaveRNN are the deep neural networks proposed to directly generate audio waveform in the field of speech synthesis[5, 6]. Recently, lossy speech coding methods using WaveNet as a non-linear predictor have been reported[7]. WaveNet has the problem that it works with up to 8 bit (μ-law) signal and takes large amount of calculation. WaveRNN solves this problem. It works with up to 16 bit signal and takes smaller amount of calculation than WaveNet.

WaveRNN outputs the predicted probability distribution of the target sample using the previous sample. In signal compression, if the probability distributions of signals are known, compression can be efficiently performed by optimal entropy coding for the distributions. The optimal entopy coding can be realized by arithmetic codeing. Generally, arithmetic coding can generate an optimal code for a set of arbitrary symbols and probabilities, and the codelength becomes the valuse close to the entropy. Therefore, by training WaveRNN to optimize entropy, it is possible to obtain the predictor that directly minimizes the codelength in arithmetic coding.

We have already done the initial examination of the speech lossless coding using WaveRNN as predictor[8]. This paper describes a new scheme of speech and audio lossless coding using WaveRNN as a nonlinear predictor.

2. MPEG-4 ALS

In the MPEG-4 ALS, the ALS with LPC (linear predictive coding) is the general audio lossless coding method. In the ALS with LPC, the PARCOR coefficients and the linear prediction residual signal are transmitted and the original signal is perfectly reconstructed by using them. Furthermore, the ALS with cascaded recursive least square–least mean square (RLS–LMS) prediction is defined[9]. In ALS with RLS-LMS prediction, it is not necessary to transmit PARCOR coefficients. Compared with the above-mentioned method, this method has not been widely used since it has a large calculation amount. However, MPEG-4 ALS with RLS-LMS prediction provides higher compression ratio than MPEG-4 ALS with LPC because it does not send PARCOR coefficients. The proposed scheme has high similarity to the MPEG-4 ALS with RLS–LMS prediction, we use it as the comparison method.

The block diagram of MPEG-4 ALS encoder and decoder with the cascaded RLS–LMS predictor id shown in Figure 1. In the encoder, the prediction signal is acquired by the cascaded RLS-LMS predictor for each frame of imput signal, and the prediction residual signal is calculated from the input signal and the prediction signal. The prediction residual signal is entropy coded and transmitted as a compressed signal. In the decoder, the transmitted compressed signal is entropy decoded to obtain the residual signal. The decoder can also calculate the prediction signal in the same way as the encoder, because the cascaded RLS-LMS predictor outputs the target prediction sample by using previous samples. Therefore, the original signal is perfectly reconstructed from the residual signal and the prediction signal obtained by the cascaded RLS-LMS predictor.

3. WaveRNN

The network structure of WaveRNN is shown in Figure 2. WaveRNN consists of single recurrent neural network (RNN) layer and two full-connected neural network (NN) layer. WaveRNN uses the previous sample as input to predict the probability distribution of the target sample. The sample bits converted into coarse bits and fine bits. The coarse bits is higher 8-bits of the 16-bits sample, and fine bits is lower 8-bits of the 16-bits sample. In Figure 2, c and f at the inputs represents coarse bits and fine bits respectively. $P(c)$ and $P(f)$ at the outputs are the occurrence probability of each coarse bits and fine bits in the next sample. Eq.(1) shows the overall flow for obtaining occurrence probabilities of coarse bits and fine bits at the target time $(P(c_t), P(f_t))$ from coarse bits and fine bits at the previous time (c_{t-1}, f_{t-1}).

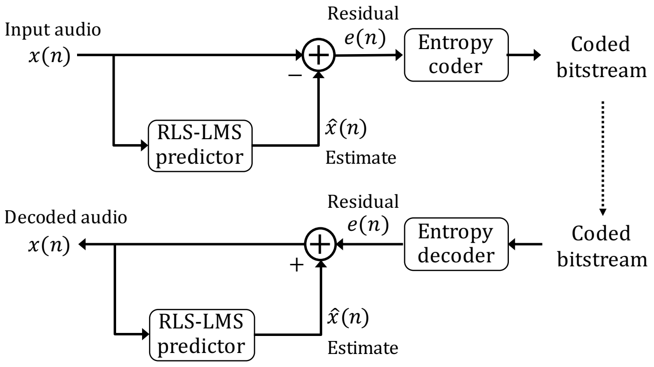

Figure 1: MPEG-4 ALS encoder (top) and decoder (bottom) with the cascaded RLS–LMS predictor.

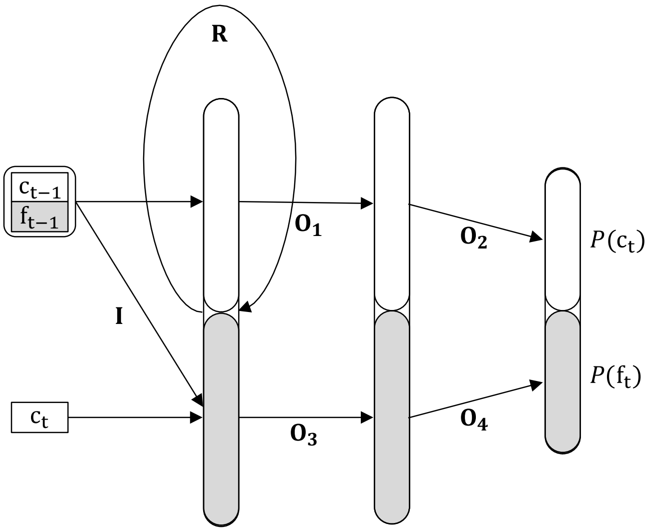

Figure 2: The network structure of WaveRNN

$$\begin{aligned} x_t &= [c_{t-1}, f_{t-1}, c_t] \\ u_t &= \sigma(R_u h_{t-1} + I_u^{\star} x_t) \\ r_t &= \sigma(R_r h_{t-1} + I_r^{\star} x_t) \\ e_t &= \tau(r_t \circ (R_e h_{t-1}) + I_e^{\star} x_t) \\ h_t &= u_t \circ h_{t-1} + (1 - u_t) \circ e_t \\ y_c, y_f &= split(h_t) \\ P(c_t) &= softmax(O_2\ relu(O_1 y_c)) \\ P(f_t) &= softmax(O_4\ relu(O_3 y_f)) \end{aligned} \tag{1}$$

Here, R_u, R_r, and R_e are the values obtained from multiplying the hidden value of the RNN layer by the recursive weight, and these are used for obtaining u_t, r_t, and e_t respectively. σ and τ are sigmoid function and hyperbolic tangent function of standard activation functions. Also, $\star$ represents a masked matrix whereby c_t is used only for computing f_t.

4. Proposed RNN-based prediction method

The block diagram of the encoder and decoder of our proposed method is shown in Figure 3. WaveRNN uses the previous sample as input to predict the probability distribution of the target sample. Arithmetic coding is performed according to combinations of the sample values and the probability distributions and it is possible to generate optimal codelengths for these combinations. WaveRNN learns to minimize the cross entropy of the true probability distribution and the predicted probability distribution. Here, the true probability distribution is that of the input, so it become one-hot vector. When perfoming arithmetic coding with the probability distribution outputs by WaveRNN, the codelength after encoding depends on the cross entropy at learning. Thus, learning in WaveRNN performs aiming at minimizing the codelength after arithmetic coding.

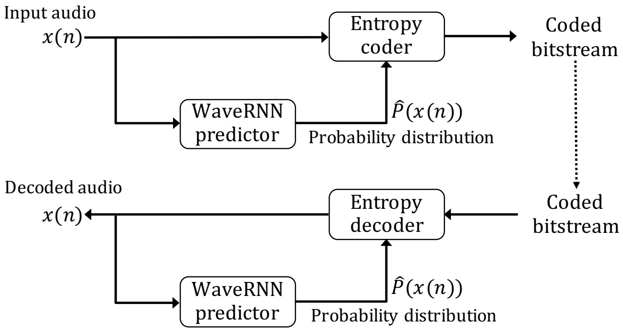

Figure 3: The encoder (top) and decoder (bottom) with WaveRNN predictor

5. Evaluation

5.1 Experimental conditions

In speech coding experiment, the speech items of Multilingual Speech Database 2002 are used[10]. Table 1 shows specifications of the speech items used in this experiment. In music coding experiments, classic music items of RWC Music Database are used[11]. Table 2 shows specifications of the music items used in this experiment. In each experiment, only some of the items were used from each database.

In speech coding experiment, the non-utterance sections are cut out from the speech items. In music coding experiment, the music items are resampled to 16 kHz and each items are split every 4 seconds. WaveRNN is proposed for speech synthesis. Therefore, it is expected that the performance will be degraded in music encoding.

Table 1: Specifications of Multilingual Speech Database 2002.

Sampling rate	16 kHz
Amplitude resolutin	16 bit
Number of channels	1 ch
Recoding time per item	4 s
Number of items	17,000

Table 2: Specifications of classic music items in RWC Music Database.

Sampling rate	44.1 kHz
Amplitude resolutin	16 bit
Number of channels	2 ch
Number of items	50
Total recoding time	5.48 hours

5.2 Criteria for performance evaluation

In this experiment, the coding performance of WaveRNN predictor is evaluated by the average codelength of coded signal. The average codelength indicate the average cross entropy per one sample of signal. This indicate the theoretical bit rate [bits/sample] in compression. The average codelength H is shown in Equation (2). Here, N is the number of samples, and $p_i(x_i)$ is the predicted probability of the true sample value x_i.

$$H = \frac{1}{N} \sum_{i=1}^{N} -\log_2 p_i(x_i) \tag{2}$$

5.3 Evaluation for speech item

In this experiment, 300 items were used for training WaveRNN and 100 items were used for evaluation. These were American English speech items by 5 men and 5 women. Figure 4 and Table 3 show the evaluation results of the performance of MPEG-4 ALS and WaveRNN predictor for speech items. In Figure 4, the error bars represent 95 % confidence intervals. WaveRNN predictor showed better compression ratio than MPEG-4 ALS in items of 95 % or more. In fact, it showed better compression ratio than MPEG-4 ALS at all items for evaluation. In the Table 3, the number in the parentheses represents the difference of average codelength from MPEG-4 ALS. The performance of WaveRNN predictor for speech item was 0.77 bits/sample better than MPEG-4 ALS in average.

Table 3: The average codelengths of speech coding with MPEG-4 ALS and WaveRNN predictor.

	WaveRNN predictor	MPEG-4 ALS
bits/sample	7.00 (-0.77)	7.77

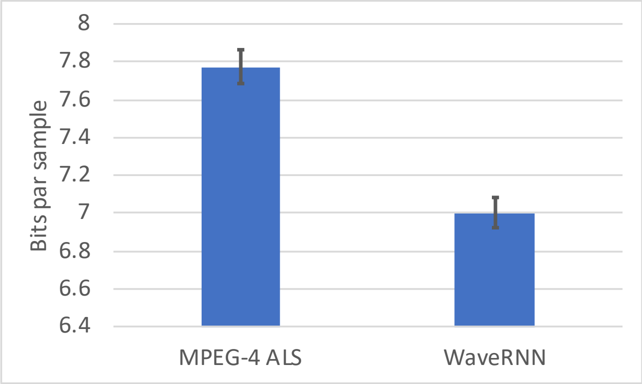

Figure 4: The evaluation results of the performance of MPEG-4 ALS and WaveRNN predictor for speech items

5.4 Evaluation for music item

In this experiment, 100 items were used for training WaveRNN and 50 items were used for evaluation. Figure 5 and Table 4 show the evaluation results of the performance of MPEG-4 ALS and WaveRNN predictor for music items. In Figure 5, the error bars represent 95 % confidence intervals. WaveRNN predictor showed worse compression ratio than MPEG-4 ALS in the average compression ratio. However, WaveRNN showed better compression ratio than MPEG-4 ALS for some sound sources (21 items in 50 items for evaluation). In the Table 4, the number in the parentheses represents the difference of average codelength from MPEG-4 ALS in average. The performance of WaveRNN predictor for music items was 0.64 bits/sample worse than MPEG-4 ALS in average.

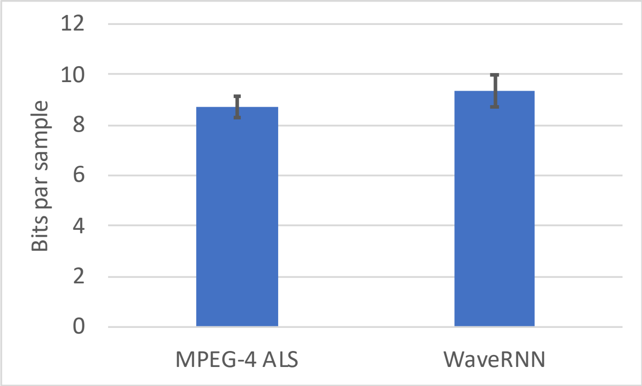

Figure 5: The evaluation results of the performance of MPEG-4 ALS and WaveRNN predictor for music items

6. Conclusions

This paper presented a new scheme of speech and audio lossless coding using WaveRNN as a predictor. In speech coding, experimental evaluation confirmed that the proposed method reduced around 0.8 bits per sample compared to MPEG-4 ALS. In audio coding, the proposed method showed worse compression ratio than MPEG-4 ALS in the average compression ratio. This is because WaveRNN used in this experiment was adjusted for speech synthesis. However, the proposed method showed better compression ratio than MPEG-4 ALS for some sound sources (21 items in 50 items for evaluation). The proposed method has the potential to obtain better compression performance than MPEG-4 ALS in music coding. It is considered that the compression performance will be improved by tuning parameters of WaveRNN for music audio. DNN-based speech and audio coding techniques will transcend the international standard technologies.

Table 4: The average codelengths of music coding with MPEG-4 ALS and WaveRNN predictor.

	WaveRNN	MPEG-4 ALS
bits/sample	9.38 (+0.64)	8.74

References

[1] ISO/IEC 14496-3:2009/Amd 3:2015, 2009 Information technology – Coding of audio-visual objects – Part 3: Audio.

[2] T. Liebechen, *et al.* ”The MPEG-4 audio lossless coding (ALS) standard - Technology and applications,” in *Proc. 119th AES Convention*, pp. 1-14, Oct, 2005.

[3] T. Liebchen and Y. Reznik, ”MPEG-4 ALS : an emerging standard for lossless audio coding,” in *Proc. IEEE Data Compression Conference*, pp. 439-448, 2004.

[4] S. Amada, *et al.* ”Experimental evaluation of encoding parameters of MPEG-4 ALS for high-resolution audio,” in *Proc. IEEE Global Conference on Consumer Electronics*, Oct, 2017.

[5] A. van den Oord, *et al.* ”WAVENET: A generative model for raw audio,” arXiv:1609.03499, 2016.

[6] N. Kalchbrenner, *et al.* ”Efficient neural audio synthesis,” arXiv:1802.08435, 2018.

[7] W. B. Kleijn, *et al.* ”WAVENET based low rate speech coding,” arXiv:1712.01120, 2017.

[8] S. Amada, *et al.* ”Experimental evaluation of the audio lossless coding with WaveRNN,” in *Proc. Acoustical Society of Japan 2018 Autumn Meeting*, 2-4-9, Sept, 2018 (in Japanese).

[9] H. Huang, *et al.* ”Cascaded RLS-LMS prediction in MPEG-4 lossless audio coding,” in *Proc. ICASSP*, pp. 181-184, May, 2006.

[10] Multilingual Speech Database 2002,
http://www.ntt-at.com/product/speech2002/

[11] RWC Music Database,
https://staff.aist.go.jp/m.goto/RWC-MDB/