WAVENET-BASED ZERO-DELAY LOSSLESS SPEECH CODING

Takenori Yoshimura, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda

Nagoya Institute of Technology
Department of Scientific and Engineering Simulation
Gokiso-cho, Showa-ku, Nagoya, Japan

ABSTRACT

This paper presents a WaveNet-based zero-delay lossless speech coding technique for high-quality communications. The WaveNet generative model, which is a state-of-the-art model for neural-network-based speech waveform synthesis, is used in both the encoder and decoder. In the encoder, discrete speech signals are losslessly compressed using sample-by-sample entropy coding. The decoder fully reconstructs the original speech signals from the compressed signals without algorithmic delay. Experimental results show that the proposed coding technique can transmit speech audio waveforms with 50% their original bit rate and the WaveNet-based speech coder remains effective for unknown speakers.

Index Terms— Speech coding, entropy coding, lossless coding, WaveNet, adaptive mel-cepstral analysis

1. INTRODUCTION

Speech coding technologies have been investigated for over 50 years and have been applied to various communication systems including broadcasting and telecommunications [1]. For two-way communications such as telephony, lossy speech coding techniques are widely used as they can achieve a very high compression ratio. However, the speech quality is invariably degraded when a sufficient bit rate is not available, and thus may not be acceptable for high-quality live communications. In the future, when communication resources are sufficiently high, lossless coding techniques [2] will be used more often. Note that lossless compression is also useful for digital music distribution over the Internet because some consumers would like to get the best possible quality of audio recordings. Although deep learning is a promising approach to achieving a high compression ratio, and thus attracted much attention in many research areas, there have been only a few attempts to use deep neural networks for speech coding [3, 4].

Several neural-network-based models for audio waveform modeling have recently been proposed. The most promising models are the WaveNet generative model [5], the SampleRNN audio generation model [6], and the WaveRNN model [7]. They use a specially designed convolutional or recurrent neural network (RNN) to capture the long-term temporal dependencies in speech signals. The WaveNet model is able to model raw audio waveforms and outperformed the best text-to-speech systems in subjective evaluation tests [5]. In addition, it has been used for not only text-to-speech but also voice conversion [8], speech denoising [9], and speech enhancement [10]. This means that this kind of model can appropriately model human speech signals. Neural-network-based nonlinear models should thus be able to compress audio speech signals more effectively than conventional coding schemes based on linear prediction [1, 11].

In this paper, we present a WaveNet-based zero-delay lossless speech coding technique. Although the concept of WaveNet-based lossless speech coding has been presented [3], it is based on the assumption of μ-law coding [12], not fully lossless speech coding. Furthermore, it suffers algorithmic delay due to computing side information such as pitch and line spectral pairs [13] depending on future inputs. Such delay interferes with real-time interactive communication. We thus investigated two types of WaveNet models: an unconditional WaveNet for which the input is only past speech samples and a conditional WaveNet for which the inputs are past speech samples and an acoustic feature calculated from those samples. Mel-cepstral coefficients, which are widely used in speech processing, are used as the acoustic feature. They are effectively calculated for every sample by using adaptive mel-cepstral analysis [14].

The contributions of this paper are as follows:

- A WaveNet-based *zero-delay lossless* speech coding technique is proposed.
- The effectiveness of the proposed technique is demonstrated in fully lossless compression, i.e., 16-bit speech coding, as well as 8-bit coding.
- The compression performance is compared with that of commonly used lossless compression techniques.

The remainder of the paper is organized as follows: Section 2 briefly explains the WaveNet generative model. The proposed WaveNet-based lossless speech coding is presented in Section 3. Section 4 describes the evaluation and discusses the results. Section 5 summarizes the key points and mentions future work.

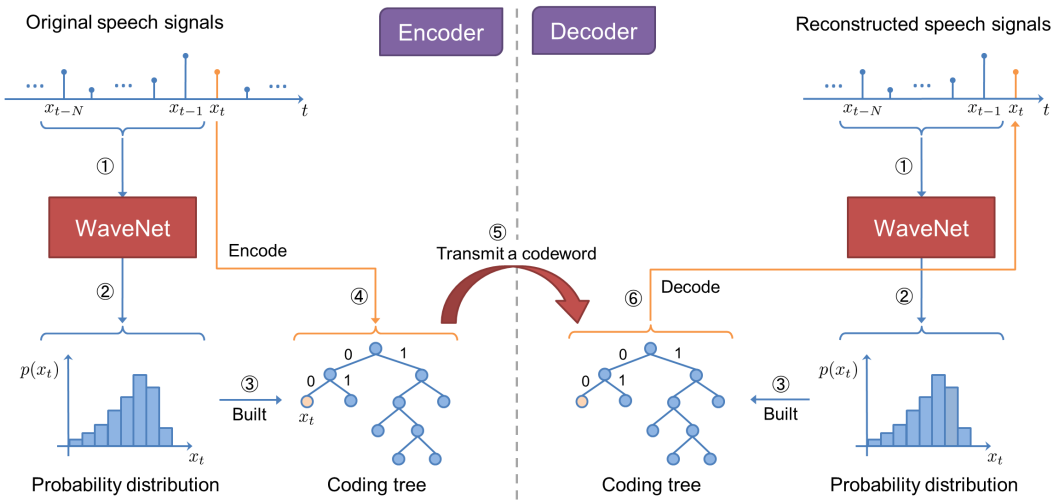

Fig. 1. Overview of proposed WaveNet-based zero-delay lossless speech coding system.

2. WAVENET GENERATIVE MODEL

WaveNet [5, 15] is an autoregressive generative model that predicts the current value of a discrete-valued time series $\boldsymbol{x} = [\, x_1\, x_2 \cdots x_T \,]$ using past samples:

$$p\left(\boldsymbol{x}\right) = \prod_{t=1}^{T} p\left(x_t \,|\, x_1, \ldots, x_{t-1}\right), \tag{1}$$

where T is the length of the time series. The conditional probability distribution in (1) is represented using a dilated causal convolution with a very large receptive field. The conditional probability distribution is categorical, resulting in flexible raw audio waveform modeling.

Given an additional input $\boldsymbol{h} = [\, \boldsymbol{h}_1\, \boldsymbol{h}_2 \cdots \boldsymbol{h}_T \,]$, WaveNet can model the conditional distribution $p\left(\boldsymbol{x} \,|\, \boldsymbol{h}\right)$. In text-to-speech synthesis, the auxiliary input $\boldsymbol{h}$ is used to produce waveforms with specified characteristics, e.g., phoneme and speaker identity. Here we assume that

$$p\left(\boldsymbol{x} \,|\, \boldsymbol{h}\right) = \prod_{t=1}^{T} p\left(x_t \,|\, x_1, \ldots, x_{t-1}, \boldsymbol{h}_1, \ldots, \boldsymbol{h}_{t-1}\right). \tag{2}$$

For more details, see [5].

3. WAVENET-BASED LOSSLESS SPEECH CODING

3.1. Overview

Figure 1 shows an overview of the proposed WaveNet-based lossless speech coding system. The basic idea of the proposed technique is sample-by-sample Huffman coding [16, 17] using a raw audio waveform generative model. That is, binary codewords are dynamically changed in accordance with the discrete probability distribution predicted by WaveNet. Since the encoder and decoder use the same WaveNet model, they can communicate using the same binary codewords. Thanks to the Huffman codes being uniquely and instantaneously decodable, sample-by-sample coding is supported.

The proposed technique comprises eight steps:

1. Set t to 1.

—Encoder—

2. Feed original speech waveform samples $x_{t-N}, \ldots, x_{t-1}$ into trained WaveNet model, where N is the receptive field size of WaveNet.
3. Build a Huffman tree from the probability distribution given by the WaveNet model.
4. Encode x_t into a codeword using the built tree and send it to the decoder.

—Decoder—

5. Feed reconstructed speech samples $x_{t-N}, \ldots, x_{t-1}$ into the WaveNet model which is the same as used in the encoder.
6. Build a Huffman tree from the probability distribution given by the WaveNet model. The built tree should be same as in the encoder.
7. Receive a codeword from the encoder and decode it into x_t using the built tree.
8. Set t to $t+1$ and go to Step 2 until $t = T$.

In this procedure, x_t $(t \leq 0)$ is zero. Auxiliary input $\boldsymbol{h}$ is calculated in Steps 2 and 5 using past samples $x_1, \ldots, x_{t-1}$, and is fed to the trained model if needed. There is no thus algorithmic delay.

3.2. Adaptive mel-cepstral analysis

In this paper, the mel-cepstral coefficients used as the auxiliary input of WaveNet. Since the standard mel-cepstral analysis [18] requires future samples, the adaptive mel-cepstral analysis [14] is used instead. The adaptive mel-cepstral analysis can recursively calculate mel-cepstral coefficients without future samples. Thus, the caclulated coefficients do not need to be transmitted: the decoder can duplicate the encoder's operation without them.

In the mel-cepstral analysis [18], the following spectral envelope model $H(z)$ is assumed to use M-th order mel-cepstral coefficients $\{\tilde{c}(m)\}_{m=0}^{M}$:

$$H(z) = \exp \sum_{m=0}^{M} \tilde{c}(m)\, \tilde{z}^{-m}, \tag{3}$$

where

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}. \tag{4}$$

The phase characteristic of the all-pass function $\tilde{z}^{-1}$ can be made to approximate the mel-frequency scale by tuning α in (4). For a given speech waveform, the mel-cepstral coefficients can be derived by minimizing $\varepsilon = E\left[e_t^2\right]$, where e_t is the output signal of the inverse filter $1/D(z)$ driven by x_t. The $D(z)$ is a minimum phase transfer function derived from $H(z) = K \cdot D(z)$ where K is the gain factor. The $H(z)$ can be decomposed in one of the ways proposed by Imai [19]:

$$K = \exp b(0), \tag{5}$$

$$D(z) = \exp \sum_{m=1}^{M} b(m)\, \Phi_m(z), \tag{6}$$

where

$$b(m) = \begin{cases} \tilde{c}(M), & (m = M) \\ \tilde{c}(m) - \alpha b(m+1), & (m < M) \end{cases} \tag{7}$$

$$\Phi_m(z) = \begin{cases} 1, & (m = 0) \\ \dfrac{\left(1 - \alpha^2\right) z^{-1}}{1 - \alpha z^{-1}} \tilde{z}^{-(m-1)}. & (m > 0) \end{cases} \tag{8}$$

Since ε is convex with respect to $\{b(m)\}_{m=1}^{M}$, the mel-cepstral coefficients can be obtained by minimizing ε with respect to $\boldsymbol{b} = [\, b(1)\, b(2) \cdots b(M)\,]^{\mathsf{T}}$ using a gradient descent method:

$$\boldsymbol{b}_{t+1} = \boldsymbol{b}_t - \mu_t \hat{\nabla} \varepsilon_t, \tag{9}$$

where $\boldsymbol{b}_t$ is the coefficient vector of $\boldsymbol{b}$ at time t and μ_t is the adaptation step size given by

$$\mu_t = \frac{k}{M \varepsilon_t}, \quad (0 < k < 1) \tag{10}$$

where ε_t is the estimate of ε at time t:

$$\varepsilon_t = \lambda\, \varepsilon_{t-1} + (1 - \lambda) e_t^2. \quad (0 \le \lambda < 1) \tag{11}$$

The $\hat{\nabla} \varepsilon_t$ is the instantaneous estimate of gradient

$$\hat{\nabla} \varepsilon_t = \tau \hat{\nabla} \varepsilon_{t-1} - 2\,(1 - \tau)\, e_t \boldsymbol{\delta}_t, \quad (0 \le \tau < 1) \tag{12}$$

$$\boldsymbol{\delta}_t = [\, \delta_t(1)\, \delta_t(2) \cdots \delta_t(M)\,]^{\mathsf{T}}, \tag{13}$$

where $\delta_t(m)$ is the output of $\Phi_m(z)$ driven by e_t. The details of the algorithm are available in [14, 20]. The derived $\boldsymbol{b}_t$ with $b_t(0) = 1/2 \log \varepsilon_t$ is converted into mel-cepstral coefficients and then used as $\boldsymbol{h}_t$ in (2).

3.3. Temperature parameter

Since WaveNet has a powerful ability to model complex speech signals, it may overfit to the training data. To avoid overfitting, we introduce a temperature parameter, $\kappa \ge 1$, to the conditional distribution in (1):

$$p\,(\boldsymbol{x}) = \prod_{t=1}^{T} \frac{1}{Z_t} p^{\frac{1}{\kappa}}\,(x_t \,|\, x_1, \ldots, x_{t-1})\,, \tag{14}$$

where Z_t is a normalization constant used to make the sum of the probabilities equal to one. The conditional distribution tends toward a uniform distribution as κ is increased, i.e., the maximum code length is expected to be small. The temperature parameter is not used for training, only for encoding and decoding.

4. EVALUATION

4.1. Experimental setup

The Carnegie Mellon University (CMU) Arctic databases [21] were used to evaluate the proposed speech coding technique. We trained a speaker-dependent WaveNet model using 22 minutes of speech data from a female speaker (*slt*). Forty sentences not including training sentences were used for evaluation. The speech signals were downsampled to 16 kHz. The dilations of the WaveNet model were set to $1, 2, 4, \ldots, 512$ [5]. Ten dilation layers were stacked three times, resulting in a receptive field with a size of 3072. The parameters were optimized using the Adam solver [22] with an initial learning rate of 0.001. The average number of bits per sample of coded speeech and the following compression ratio r were used as objective measures:

$$r = \frac{\text{Original size}}{\text{Compressed size}}. \tag{15}$$

4.2. 8-bit speech coding

The downsampled speech signals were quantized to eight bits and were modeled using the unconditional WaveNet model,

Table 1. Performance comparison for 8-bit speech coding.

	compression ratio	bits/sample
Original	1.00	8.00
GNU Gzip	1.24	6.43
shorten	1.59	5.02
MPEG-4 ALS	1.70	4.71
WaveNet	**2.68**	**2.98**

Table 2. Performance comparison for 16-bit speech coding.

	compression ratio	bits/sample
Original	1.00	16.0
GNU Gzip	1.19	13.4
shorten	1.70	9.40
MPEG-4 ALS	1.83	8.75
WaveNet	**2.24**	**7.13**

which had a channel size for dilations, residual blocks, and skip-connections of 32 [5]. There were 1,000,000 training steps. Since a μ-law quantizer is widely used in 8-bit case, we modeled μ-law speech instead of 8-bit PCM. The temperature parameter κ was set to 1.0. We compared the performance of the proposed technique with that of several commonly used coding schemes: GNU Gzip[1], shorten[2] [23], and MPEG-4 ALS[3] [24, 25]. Although these schemes can also be used for audio coding, we evaluated them in terms of speech coding ability. As shown in Table 1, the proposed WaveNet-based coding technique achieved considerably better compression than the well-known compression tools. The quantization from original 16-bit to 8-bit facilitates the optimization of the model parameters of WaveNet at the expense of speech quality. One way to improve speech quality is to use noise shaping [26].

4.3. 16-bit speech coding

The downsampled 16-bit speech signals were modeled using the unconditional WaveNet model with a softmax output layer. The size of the channel for dilations, residual blocks, and skip-connections was 64. The temperature parameter κ was set to 1.2. As shown in Table 2, the proposed technique again outperformed the commonly used coding schemes. The average number of bits per sample was below eight; i.e., the compressed size was less than half of original one. Since the computational time for speech coding was very long due to the generation of a Huffman tree with a size of 2^{16} for every sample, we investigated the performance of the proposed technique for 8-bit speech coding in subsequent experiments.

4.4. Speech coding with auxiliary input

To investigate the effectiveness of auxiliary input calculated from only past speech samples, a conditional WaveNet model was trained using 500,000 training steps. The auxiliary input consisted of 24-th order mel-cepstral coefficients including the zeroth coefficient extracted using adaptive mel-cepstral analysis[4]. The tuning parameters for the algorithm were set to $\alpha = 0.42$, $\lambda = 0.99$, $\tau = 0.95$, and $k = 0.1$. As shown in Table 3, the conditional WaveNet obtained slightly better results than the unconditional one. The auxiliary input helped to speed up model training and improved compression performance to some extent.

[1]https://www.gnu.org/software/gzip/
[2]http://www.etree.org/shnutils/shorten/
[3]http://www.nue.tu-berlin.de/mp4als
[4]http://sp-tk.sourceforge.net/

Table 3. Performance comparison between unconditional and conditional WaveNet models.

	compression ratio	bits/sample
WaveNet (uncond.)	2.68	2.98
WaveNet (cond.)	**2.72**	**2.94**

4.5. Speech coding for unknown speakers

We also investigated the speaker dependency of the proposed WaveNet-based lossless speech coding. Three male speakers (*bdl*, *jmk*, and *rms*) and a female speaker (*clb*) in the CMU databases and the same 40 sentences were used for evaluation. Each of the sentences was fed into the WaveNet model of the training speaker (*slt*), a female speaker. As shown in Figure 2, although the compression ratios for two of the male speakers, *bdl* and *rms*, were much lower than that of the training speaker, they were still higher than 2.0. In comparison with the ratios for the unconditional WaveNet, those for the conditional WaveNet were relatively high. This indicates that using the auxiliary input is effective for unknown speakers as well as for the training speaker.

5. CONCLUSION

A WaveNet-based zero-delay lossless speech coding technique was presented. Objective experiments showed that it outperformed widely used speech coding techniques. Future work includes investigating the effectiveness of a speaker-independent WaveNet-based speech coder and compressing musical audio signals using the proposed technique. Using a mixture of logistic distributions [15] as the output layer of the WaveNet model is also future work.

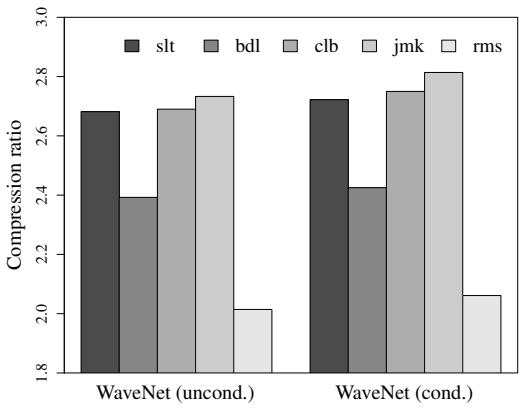

Fig. 2. Compression ratios for five speakers in CMU Arctic databases.

6. ACKNOWLEDGMENT

This work was partly supported by JSPS KAKENHI Grant Number JP18K11163 and CASIO Science Promotion Foundation.

7. REFERENCES

[1] T. Moriya, R. Sugiura, Y. Kamamoto, H. Kameoka, and N. Harada, "Progress in LPC-based frequency-domain audio coding," *APSIPA Transactions on Signal and Information Processing*, vol. 5, pp. 1–10, 2016.

[2] M. Hans and R. W. Schafer, "Lossless compression of digital audio," *IEEE Signal Processing Magazine*, vol. 5, pp. 21–32, 2001.

[3] W. B. Kleijn et al., "WaveNet based low rate speech coding," *arXiv:1712.01120*, 2017.

[4] Srihari Kankanahalli, "End-to-end optimized speech coding with deep neural networks," *Proc. of ICASSP*, pp. 2521–2525, 2018.

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

[6] S. Mehri et al., "SampleRNN: An unconditional end-to-end neural audio generation model," *arXiv:1612.07837*, 2016.

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

[8] J. Niwa, T. Yoshimura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Statistical voice conversion based on WaveNet," *Proc. of ICASSP*, pp. 5289–5293, 2018.

[9] D. Rethage, J. Pons, and X. Serra, "A WaveNet for speech denoising," *arXiv:1706.07162*, 2017.

[10] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florencio, and M. Hasegawa-Johnson, "Speech enhancement using Bayesian WaveNet," *Proc. of Interspeech*, pp. 2013–2017, 2017.

[11] P. P. Vaidyanathan, *The Theory of Linear Prediction*, Morgan & Claypool Publishers, 2007.

[12] "Pulse code modulation (PCM) of voice frequencies," *ITU-T Recommendation G.711*, 1988.

[13] D. Rowe, "Codec 2 - open source speech coding at 2400 bits/s and below," *TAPR and ARRL 30th Digital Communications Conference*, pp. 80–84, 2011.

[14] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," *Proc. of ICASSP*, vol. 1, pp. 137–140, 1992.

[15] A. van den Oord et al., "Parallel WaveNet: Fast high-fidelity speech synthesis," *arXiv:1711.10433*, 2017.

[16] D. A. Huffman, "A method for the construction of minimum-redundancy codes," *Proc. of Institute of Radio Engineering*, vol. 40, pp. 1098–1101, 1952.

[17] J. Schmidhuber and S. Heil, "Sequential neural text compression," *IEEE Transactions on Neural Networks*, vol. 7, no. 1, pp. 142–146, 1996.

[18] K. Tokuda, T. Kobayashi, T. Chiba, and S. Imai, "Spectral estimation of speech by mel-generalized cepstral analysis," *Electronics and Communications in Japan (Part 3)*, vol. 76, no. 2, pp. 30–43, 1993.

[19] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (MLSA) filter for speech synthesis," *Electronics and Communications in Japan*, vol. 66, no. 2, pp. 11–18, 1983.

[20] K. Tokuda, H. Matsumura, T. Kobayashi, and S. Imai, "Speech coding based on adaptive mel-cepstral analysis," *Proc. of ICASSP*, vol. 1, pp. 197–200, 1994.

[21] J. Kominek and A. W. Black, "CMU ARCTIC databases for speech synthesis," *Technical Report CMU-LTI-03-177*, pp. 1–19, 2003.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.

[23] T. Robinson, "SHORTEN: Simple lossless and near-lossless waveform compression," *Technical report CUED/F-INFENG/TR.156*, 1994.

[24] T. Liebchen and Y. A. Reznik, "MPEG-4 ALS: An emerging standard for lossless audio coding," *Proc. of Data Compression Conference*, pp. 439–448, 2004.

[25] T. Liebchen, T. Moriya, N. Harada, Y. Kamamoto, and Y. A. Reznik, "The MPEG-4 audio lossless coding (ALS) standard - technology and application," *Proc. of AES 119th Convention*, 2005.

[26] T. Yoshimura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Mel-cepstrum-based quantization noise shaping applied to neural-network-based speech waveform synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 7, pp. 1173–1180, 2018.