SPEAKER-DEPENDENT WAVENET-BASED DELAY-FREE ADPCM SPEECH CODING

Takenori Yoshimura[⋆†]*, Kei Hashimoto*[⋆]*, Keiichiro Oura*[⋆]*, Yoshihiko Nankaku*[⋆]*, and Keiichi Tokuda*[⋆]

[⋆] Nagoya Institute of Technology, Department of Computer Science, Nagoya, Japan
[†] Nagoya University, Institute of Innovation for Future Society, Nagoya, Japan

ABSTRACT

This paper proposes a WaveNet-based delay-free adaptive differential pulse code modulation (ADPCM) speech coding system. The WaveNet generative model, which is a state-of-the-art model for neural-network-based speech waveform synthesis, is used as the adaptive predictor in ADPCM. To further improve speech quality, mel-cepstrum-based noise shaping and postfiltering were integrated with the proposed ADPCM system. Both objective and subjective evaluation results indicate that the proposed ADPCM system outperformed not only the conventional ADPCM system based on ITU-T Recommendation G.726 but also the ADPCM system based on adaptive mel-cepstral analysis.

Index Terms— Speech coding, ADPCM, WaveNet, mel-cepstrum, noise shaping

1. INTRODUCTION

Adaptive differential pulse code modulation (ADPCM) [1] is a lossy waveform coding technique widely used in telecommunications. The idea behind ADPCM is based on the fact that a speech sample can be roughly predicted from its past samples. Instead of quantizing and transmitting a speech sample directly, ADPCM quantizes and transmits the difference between the actual speech sample and a predicted one. Since the quantization step size dynamically changes according to the behavior of the speech samples, ADPCM can effectively transmit waveforms with reasonable speech quality.

The most widely used standardized ADPCM system was provided in ITU-T Recommendation G.726 [2]. It includes a backward adaptive quantizer and adaptive predictor with two structures: a sixth-order section that models zeros and second-order section that models poles of a transfer function from input speech samples. Tokuda et al. [3] proposed a short-term adaptive predictor based on adaptive mel-cepstral analysis [4]. There have been other attempts to improve the adaptive predictor in ADPCM using a simple structure [5, 6]. However, neural-network-based nonlinear models should reconstruct speech samples more precisely than such linear predictive coders. In the 1990s, ADPCM systems using a neural-network-based nonlinear predictor were proposed [7, 8, 9, 10]. Although their effectiveness was shown through experiments, the neural network structures were very simple, e.g., only one hidden layer with tens of neurons. Recent neural-network-based waveform models, such as WaveNet [11] and SampleRNN [12], are thus expected to more accurately predict speech samples. Since the WaveNet generative model can effectively handle long time-range signals due to dilated causal convolutions [11], it should work as not only a short-term but also long-term predictor of speech samples.

As the effectiveness of WaveNet has been proven in various domains, e.g., text-to-speech synthesis [11, 13], parametric speech coding [14] and lossless speech coding [14, 15], we propose a WaveNet-based ADPCM speech coding system in which WaveNet is used as the adaptive predictor. Since the WaveNet model in the proposed system is unconditional, there is neither computational cost nor algorithmic delay incurred by calculating side information such as line spectral pairs, i.e., delay is at most one sample if computational speed is fast enough. Inspired from previous work [3], mel-cepstrum-based noise shaping and postfiltering were integrated with the proposed system to improve the quality of reconstructed speech. The spectrum represented by mel-cepstral coefficients has a frequency resolution similar to that of the human ear. Since the transfer functions of noise shaping and postfiltering are defined through the mel-cepstral coefficients, the effects of noise shaping and postfiltering should be suitable to human auditory perception.

The remainder of the paper is organized as follows: Section 2 briefly explains the basic structure of ADPCM. The proposed WaveNet-based ADPCM speech coding system with mel-cepstrum-based noise shaping and postfiltering is presented in Section 3. Section 4 describes the objective and subjective evaluations, and discusses the results. Finally, conclusion and future work are mentioned in Section 5.

2. BASIC ADPCM STRUCTURE

Figure 1 shows the basic structure of an ADPCM coder. Given an input sample $x[n]$, the difference in that and a predicted one is calculated. The difference signal $e[n]$ is quantized using an adaptive quantizer. Then, the quantized difference signal, $i[n]$, is transmitted to a decoder through a digital channel. The decoder generates a reconstructed sample $\hat{x}[n]$ from the received $i[n]$. In the figure, the gain

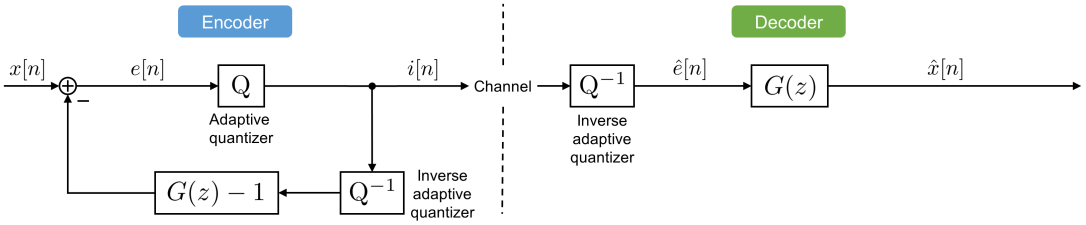

Fig. 1. Basic structure of ADPCM coder.

factor of a filter $G(z)$ is assumed to be unity, i.e., the impulse response at time $n = 0$ is unity. Thus, $G(z) - 1$ has no delay-free paths.

3. WAVENET-BASED ADPCM SPEECH CODING

3.1. WaveNet generative model

WaveNet [11] is an autoregressive generative model that predicts the current sample of a discrete-valued time series $\boldsymbol{x} = [\, x[0]\, x[1]\, \cdots\, x[N-1]\,]$ using past samples:

$$p\left(\boldsymbol{x}\right) = \prod_{n=0}^{N-1} p\left(x[n] \,|\, x[0], x[1], \ldots, x[n-1]\right), \qquad (1)$$

where N is the length of the time series. The conditional probability distribution in (1) is represented using dilated causal convolutions with a very large receptive field. The output of the WaveNet model is a categorical distribution.

3.2. WaveNet-based adaptive predictor

The adaptive filter $G(z)$ based on the WaveNet model should accurately predict speech samples, resulting in improved quality of decoded speech. The relationship between a past sample $x[n-1]$ and the predicted next sample $x'[n]$ can be represented using the WaveNet model $P(z)$ (see Fig. 2(a)). The relationship can be converted as the relationship between $x[n]$ and $x'[n]$ using a delay operator z^{-1} (Fig. 2(b)). The relationship and inverse relationship between the difference $e[n] = x'[n] - x[n]$ and $x[n]$ are shown in Figs. 2(c) and 2(d), respectively. It can be seen from these figures that

$$G(z) = \frac{X(z)}{E(z)} = \frac{1}{1 - z^{-1}P(z)}. \qquad (2)$$

Thus,

$$G(z) - 1 = \frac{z^{-1}P(z)}{1 - z^{-1}P(z)}. \qquad (3)$$

From (2) and (3), the block diagram of the proposed WaveNet-based ADPCM speech coding system is derived as that shown in Fig. 3.

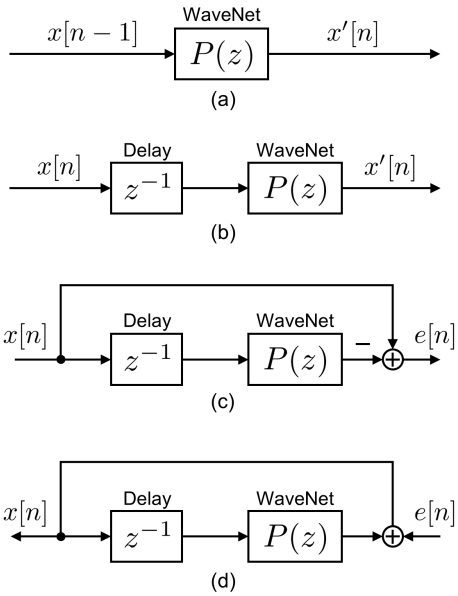

Fig. 2. Relationship between two signals.

Since the WaveNet model outputs a distribution rather than a scalar value, the output must be converted to a scalar value $x'[n]$ to calculate the difference signal $e[n]$. In this paper, a weighted trimmed mean is used for the conversion using only reliable values in the distribution:

$$x'[n] = \frac{1}{B} \sum_{j \in J_B^{(n)}} p_j^{(n)} \cdot x_j, \qquad (4)$$

where $J_B^{(n)}$ is a set of indices that have B highest probabilities in the predicted distribution $p^{(n)}$, $p_j^{(n)}$ is a probability corresponding to the j-th bin of $p^{(n)}$, and x_j is a sample value corresponding to the j-th bin of the distribution. If $B = 1$, only a sample value corresponding to the mode of the predicted distribution is used as a predicted sample.

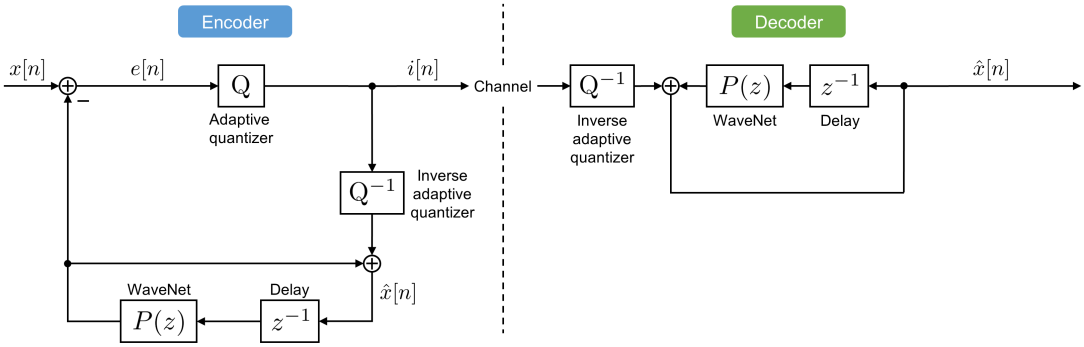

Fig. 3. Block diagram of proposed WaveNet-based ADPCM system.

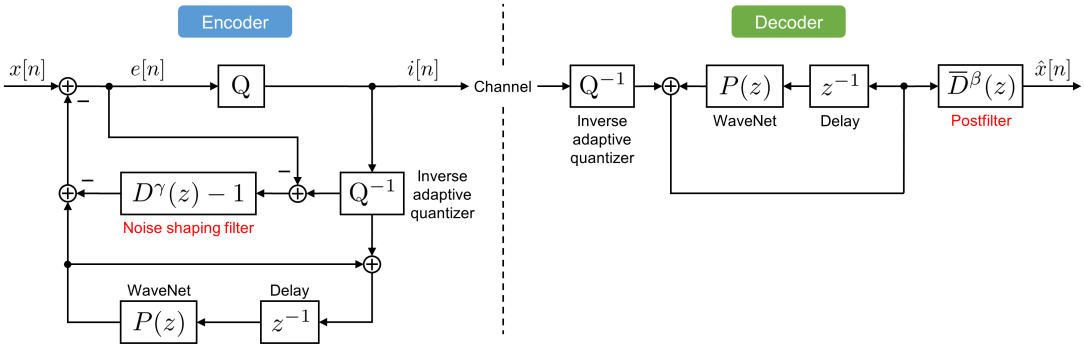

Fig. 4. Block diagram of proposed WaveNet-based ADPCM system with noise shaping and postfiltering.

3.3. Mel-cepstrum-based noise shaping and postfiltering

Figure 4 shows a block diagram of the proposed WaveNet-based ADPCM system with mel-cepstrum-based noise shaping and postfiltering. In the figure, $D(z)$ is a minimum phase transfer function derived from the following spectral envelope model using M-th order mel-cepstral coefficients $\{\tilde{c}(m)\}_{m=0}^{M}$, which are calculated from input speech samples by using adaptive mel-cepstral analysis [4]:

$$\begin{aligned} H(z) &= \exp \sum_{m=0}^{M} \tilde{c}(m)\, \tilde{z}^{-m} \\ &= K \cdot D(z), \end{aligned} \tag{5}$$

where $\tilde{z}^{-1} = (z^{-1} - \alpha)/(1 - \alpha z^{-1})$ and K is the gain factor. The phase characteristic of the all-pass function $\tilde{z}^{-1}$ can approximate the mel-frequency scale by tuning α. By following one of the methods proposed in [16], we can derive

$$K = \exp\, b(0), \tag{6}$$

$$D(z) = \exp \sum_{m=1}^{M} b(m)\, \Phi_m(z), \tag{7}$$

where

$$b(m) = \begin{cases} \tilde{c}(M), & (m = M) \\ \tilde{c}(m) - \alpha b(m+1), & (m < M) \end{cases} \tag{8}$$

$$\Phi_m(z) = \begin{cases} \dfrac{(1-\alpha^2)z^{-1}}{1-\alpha z^{-1}} \tilde{z}^{-(m-1)}, & (m > 0) \\ 1. & (m = 0) \end{cases} \tag{9}$$

The postfiltering filter $\overline{D}(z)$ is represented as

$$\overline{D}(z) = \exp \sum_{m=1}^{M} \bar{b}(m)\, \Phi_m(z), \tag{10}$$

where

$$\bar{b}(m) = \begin{cases} b(m), & (m > 1) \\ -\alpha b(2). & (m = 1) \end{cases} \tag{11}$$

The tunable parameters γ and β control the effects of noise shaping and postfiltering, respectively. These filters can be implemented using a structure of mel-log spectrum approximatation (MLSA) filter [16].

4. EXPRIMENTS

4.1. Experimental setup

We used the Carnegie Mellon University (CMU) Arctic databases [17] to evaluate the proposed system. The speech signals in the databases were downsampled to 8 kHz and quantized to eight bits by using the μ-law quantizer [18]. A speaker-dependent WaveNet model was trained using 1092 sentences uttered by a female speaker (*slt*). Forty sentences not including training sentences were used for evaluation. The dilations of the WaveNet model were set to $1, 2, 4, \ldots, 512$. Ten dilation layers were stacked twice. The size of the channel for dilations, residual blocks, and skip-connections was 32. The parameters were optimized through the Adam solver [19]. We compared the ADPCM methods based on G.726, adaptive mel-cepstral analysis, and WaveNet. The details of the methods are summarized in Table 1. The backward adaptive quantizer used in G.726 without slow scale factors [2] was used in all the methods. We set $\alpha = 0.31$, $\gamma = 0.3$, $\beta = 0.3$, $M = 12$, and $B = 200$.

4.2. Objective experimental results

We used the mean opinion score-listening quality objective (MOS-LQO) [20, 21] measure calculated from the perceptual evaluation of speech quality [22]. The MOS-LQO values for all method are shown in Fig. 5. As expected, the increase in bit rate drastically improved the MOS-LQO. **AMCEP2** outperformed **AMCEP**, which achieved a better score than that of **G726-16**. This indicates the effectiveness of mel-cepstrum-based noise shaping and postfiltering. Although **WN-SD** did not use them, it was comparable to **AMCEP2**. **WN-SD2** outperformed **G726-24** in terms of MOS-LQO.

4.3. Subjective experimental results

The speech quality of reconstructed signals was subjectively assesed by eight participants through the MOS test. Each participant evaluated 10 sentences randomly chosen from the 40 test sentences, i.e., each participant rated 70 sentences (10 sentences $\times$ 7 methods). The subjective experiments were conducted in a soundproof room. Figure 6 shows the results. The relative relationship among the methods was roughly similar to that of the objective evaluation. However, **WN-SD2** was comparable to **G726-32**. This indicates that the proposed method can improve a transmission efficiency by two fold compared with the widely used conventional G.726-based method.

Table 1. Methods to compare performance.

	Bit rate [kbit/s]	Type of predictor	Noise shaping & postfiltering
G726-16	16	G.726 [2]	
G726-24	24	G.726	
G726-32	32	G.726	
AMCEP	16	MLSA [3]	
AMCEP2	16	MLSA	✓
WN-SD	16	WaveNet	
WN-SD2	16	WaveNet	✓

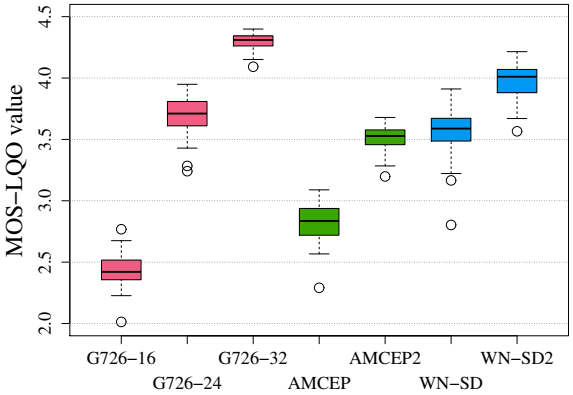

Fig. 5. Boxplot of MOS-LQO values.

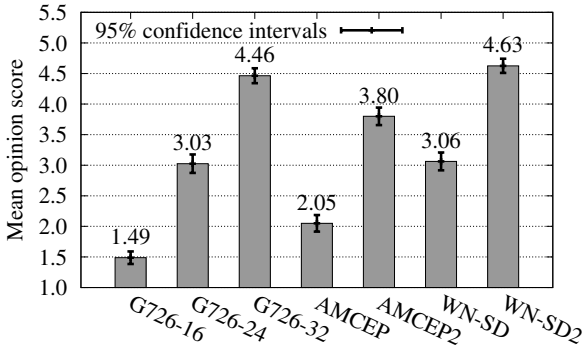

Fig. 6. Mean opinion scores for speech quality.

5. CONCLUSION

A speaker-dependent WaveNet-based delay-free ADPCM speech coding system was presented. Experiments showed that it outperformed conventional ADPCM coding systems. Future work includes investigating the speaker dependency and the robustness against bit errors of the proposed system.

6. ACKNOWLEDGMENTS

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

7. REFERENCES

[1] P. Cummiskey, N. S. Jayant, and J. L. Flanagan, "Adaptive quantization in differential PCM coding of speech," *The Bell System Technical Journal*, vol. 52, no. 7, pp. 1105–1118, 1973.

[2] "40, 32, 24, 16 kbit/s adaptive differential pulse code modulation (ADPCM)," *ITU-T Recommendation G.726*, 1990.

[3] 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.

[4] 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.

[5] E. Kruger and H. W. Strube, "Linear prediction on a warped frequency scale," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 9, pp. 1529–1531, 1988.

[6] V. Despotovic and Z. Peric, "ADPCM using a second-order switched predictor and adaptive quantizer," *Advances in Electrical and Computer Engineering*, vol. 11, no. 3, pp. 61–64, 2011.

[7] S. Haykin and L. Li, "16kb/s adaptive differential pulse code modulation of speech," *Proc. of the International Workshop on Applications of Neural Networks to Telecommunications*, pp. 132–138, 1993.

[8] S. Bartolini, F. Bartolini, V. Cappellini, and A. Mecocci, "EEG signal compression based on ADPCM and neural network predictors," *Proc. of 14th Colloque GRETSI*, pp. 1287–1290, 1993.

[9] F. Bartolini, V. Cappellini, S. Nerozzi, and A. Mecocci, "Recurrent neural network predictors for EEG signal compression," *Proc. of ICASSP*, vol. 5, pp. 3395–3398, 1995.

[10] M. Faundez-Zanuy, F. Vallverdu, and E. Monte, "Nonlinear prediction with neural nets in ADPCM," *Proc. of ICASSP*, vol. 1, pp. 345–348, 1998.

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

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

[13] Akira Tamamori, Tomoki Hayashi, Kazuhiro Kobayashi, Kazuya Takeda, and Tomoki Toda, "Speaker-dependent WaveNet vocoder," *Proc. of Interspeech*, pp. 1118–1122, 2017.

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

[15] Takenori Yoshimura, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda, "WaveNet-based zero-delay lossless speech coding," *Proc. of 2018 IEEE Workshop on Spoken Language Technology*, 2018.

[16] 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.

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

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

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

[20] "Mapping function for transforming P.862 raw result scores to MOS-LQO," *ITU-T Recommendation P.862.1*, 2003.

[21] "Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs," *ITU-T Recommendation P.862.2*, 2007.

[22] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *ITU-T Recommendation P.862*, 2001.