LOW BIT-RATE WIDEBAND SPEECH CODING: A DEEP GENERATIVE MODEL BASED APPROACH

Gang Min[1]*, Xiongwei Zhang*[2]*, Xia Zou*[2]*, Xiangyang Liu*[1]

[1]Institute of Information and Communication, National University of Defense Technology, China
[2]Army Engineering University of PLA, China

ABSTRACT

Traditional low bit-rate speech coding approach only handles narrowband speech at 8kHz, which limits further improvements in speech quality. Motivated by recent successful exploration of deep learning methods for image and speech compression, this paper presents a new approach through vector quantization (VQ) of mel-frequency cepstral coefficients (MFCCs) and using a deep generative model called WaveGlow to provide efficient and high-quality speech coding. The coding feature is sorely an 80-dimension MFCCs vector for 16kHz wideband speech, then speech coding at the bit-rate throughout 1000-2000 bit/s could be scalably implemented by applying different VQ schemes for MFCCs vector. This new deep generative network based codec works fast as the WaveGlow model abandons the sample-by-sample autoregressive mechanism. We evaluated this new approach over the multi-speaker TIMIT corpus, and experimental results demonstrate that it provides better speech quality compared with the state-of-the-art classic MELPe codec at lower bit-rate.

Index Terms— speech coding, mel-frequency cepstral coefficients, vector quantization, WaveGlow

1. INTRODUCTION

Low bit-rate speech coding, which encodes speech signals at the bit rate below 4800 bit/s, has widespread applications in the field of both satellite and secure communications. Many successful low bit-rate speech coding algorithms have been proposed in the literatures, such as linear predictive coding (LPC-10) [1], code-excited linear prediction (CELP) [2], mixed excitation linear prediction (MELP) [3], etc. However, high-quality speech coding under low bit-rate conditions still faces great challenge, especially for the wideband speech and in the presence of background acoustic noises. All of the classic speech vocoders mentioned above belong to the source-filter speech coding framework, in which the speech coding parameters include linear prediction coefficients (LPCs), pitch, energy, etc. Different types of speech coding parameters are rarely quantized together, so it is very difficult to further reduce the speech coding rate. Therefore, many other speech coding methods have been studied towards alternatives to the classic linear prediction coding model.

MFCC codec encodes speech signals through scalar quantization (SQ) or vector quantization (VQ) of MFCCs, which provides a new promising scheme for speech coding at low bit-rate conditions [4][5]. However, there are still some limitations need to be resolved for further improving the total performance. The first is the quality of coded speech needs further improvement, since there exists spectrum smearing problem, especially in the high-frequency region, which is caused by using the overlapped triangle window with mel-frequency scale for MFCCs extraction. Another is the processing efficiency also needs improvement since the traditional MFCC codec uses the Griffin-Lim algorithm (GLA) to estimate the lost phase information via discrete Fourier transform (DFT) and inverse discrete Fourier transform (IDFT) iteratively [6]. However, GLA suffers from slow convergence problem when the random initialization of the phase spectrogram is not ideal. Moreover, current MFCC codec is rarely able to handle 16kHz wideband speech signals [4].

In the last decade, deep learning methods have been used for dramatically improving the performance of many speech processing applications, such as speech enhancement (SE), text-to-speech (TTS), automatic speech recognition (ASR), etc. Most recently, deep neural networks have shown to be promising in handling the traditional speech coding task [7]. One of the most representative works is WaveNet based codec [8][9], which uses WaveNet as a generative model to synthesize speech waveforms from the bitstream generated from traditional speech codecs, such as codec2, MELP, etc. WaveNet is a kind of autoregressive neural networks-based model which generates high-quality speech waveforms, however, WaveNet suffers from very slow inference speed, which prevents its real-time speech coding applications. Besides, other models such as simple RNN, LPCNet are also explored for speech and audio coding applications [10]-[12]. The authors in [13] presented Deep Vocoder, which compresses narrowband speech with deep autoencoder and uses GLA to recover speech signals from decoded speech spectrogram, similar work in [14] presented DeepVoCoder which uses a convolutional neural network (CNN)-based encoder model to compress speech signals. However, the quality of coded

speech and the efficiency of speech decoding need further improvement for real-word communication applications.

Recent research on TTS using deep generative models conditioned on mel-spectrogram motivates our study in combining quantization of MFCCs with efficient and high-quality speech generative models for speech coding task in this paper. WaveGlow is a flow-based deep generative network, which delivers speech quality almost as good as WaveNet, however, the inference speed of WaveGlow is much faster than which of WaveNet because it abandons the sample-by-sample autoregressive mechanism [15]. Recent study of comparison on neural vocoders for speech reconstruction from mel-spectrogram also confirmed the superiority of WaveGlow for making tradeoff between speech quality and computational complexity [16]. Therefore, we choose WaveGlow as the generative model to synthesize speech waveforms from the quantized mel-spectrogram. The coding feature in our vocoder is an 80-dimension MFCCs vector for 16kHz wideband speech signal, then speech coding at the bit-rate throughout 1000-2000 bit/s could be scalably implemented with different quantization schemes of MFCCs vector.

2. ALGORITHM

2.1. Speech Coding with MFCCs and WaveGlow

Speech coding model is the basis for converting speech signals to bitstream. Like traditional speech vocoders, there are mainly three steps for speech coding with quantization of MFCCs and WaveGlow, which are extraction of speech coding features, quantization of these features and speech synthesis from quantized feature parameters, as is shown in Fig.1.

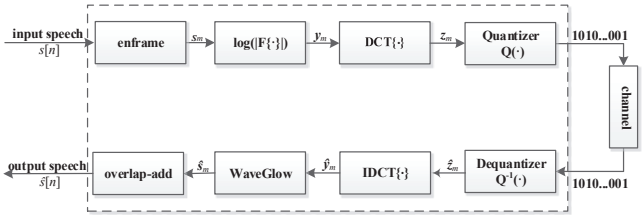

Fig. 1. Overview of the proposed vocoder.

Let *s*[*n*] denote the speech waveforms, then it is enframed by a window $w[n]$,

$$s_m[n] = s[mR + n]\, w[n] \tag{1}$$

where $L(0 \le n \le L-1)$ denotes the window length, *R* denotes the frame shift, $m(m = 1, 2, ..., M)$ denotes the frame index. At this time, each speech frame is concisely denoted as follows,

$$\boldsymbol{s}_m = [s_m(0), s_m(1), ..., s_m(L-1)]^{\mathsf{T}} \tag{2}$$

Then, the log mel-spectrogram of each speech frame can be computed as,

$$\boldsymbol{y}_m = \mathbf{M} \log(|\mathrm{F}\{\boldsymbol{s}_m\}|) \tag{3}$$

where $\mathrm{F}\{\boldsymbol{s}_m\}$ is the *N*-point fast Fourier transform (FFT) of $\boldsymbol{s}_m$, $|\cdot|$ denotes the modulus of a complex number. Due to the symmetry, the latter $N/2-1$ elements of $|\mathrm{F}\{\boldsymbol{s}_m\}|$ will be discarded. $\mathbf{M} \in \mathbb{R}^{K\times(N/2+1)}$ denotes the mel-filter weighting matrix, where K is the number of mel-filter bands.

Furthermore, the MFCCs vector of each speech frame can be computed as follows,

$$\boldsymbol{z}_m = \mathrm{DCT}\{\boldsymbol{y}_m\} \tag{4}$$

where $\mathrm{DCT}\{\cdot\}$ denotes the discrete cosine transform.

At the transmitter, the quantizer $Q(\cdot)$ uses the SQ or VQ technique to quantize MFCCs vector $\boldsymbol{z}_m$, and converts it to bitstream, which is then modulated for transmitting.

At the receiver, the quantized MFCCs vector $\hat{\boldsymbol{z}}_m$ is recovered by searching the codebook by the dequantizer $Q^{-1}(\cdot)$. Then, the reconstructed log mel-sepctrogram $\hat{\boldsymbol{y}}_m$ is computed by inverse discrete cosine transform (IDCT) of $\hat{\boldsymbol{z}}_m$, which is then used for conditioning of WaveGlow in order to synthesize speech frame. At last, the speech waveforms $\hat{s}[n]$ is reconstructed by the overlap-add operation.

2.2. Quantization of MFCCs

The quantization step of feature parameters is crucial for reducing the bit-rate of speech coding and maintaining high-quality of coded speech. Conventional speech vocoders contain different types of speech coding parameters, which are rarely quantized together. However, the speech coding parameters in the proposed vocoder mentioned above are solely MFCCs vector, so scalable speech coding schemes at different bit-rate could be implemented conveniently using the SQ or VQ technique. The first element of MFCCs vector represents energy, where its value and variance is significantly greater than other elements, so it is independently quantized using the SQ technique. As for other elements of MFCCs vector, they represent the vocal and excitation parameters, which are quantized together using the VQ technique.

2.3. Conditional WaveGlow as a Decoder

WaveGlow is a flow-based deep neural generative model for synthesizing high-quality speech signals conditioned on mel-spectrogram. Previous study has shown that Mean Opinion Score (MOS) of the synthesized speech via WaveGlow is able to reach up to 3.9 on the LJ speech corpus [15], so trying to use WaveGlow as a decoder for speech coding is very attractive. WaveGlow consists of a series of invertible flow layers that transforms a simple zero mean spherical Gaussian distribution to one which has the desired speech distribution [15]. WaveGlow network could be directly trained by minimizing

Table 1. Bit allocation scheme for MFCCs quantization.

f_s	L	R	Rate	Bits/	Quantizaiton Scheme	
(Hz)	(sample)	(sample)	(bit/s)	frame	Energy(z_0)	Formant and Pitch ($z_2 \sim z_{80}$)
16000	1024	256	1000	16	4-bit SQ	12-bit VQ
16000	1024	256	2000	32	6-bit SQ	(13-13)-bit MSVQ

the negative log-likelihood of for training set. Once the WaveGlow network is trained, doing inference to generate speech waveforms from quantized mel-spectrogram could be implemented by sampling from a Gaussian distribution and putting them through the WaveGlow network.

2.4. Bit Allocation Scheme for Speech Coding

Bit allocation is an important procedure for determining the bit-rate of speech coding. As previously discussed, the first element of MFCCs vector z_0 and other elements of MFCCs vector $z_2 \sim z_{80}$ are quantized using different methods, respectively. The proposed vocoder proceeds with the wideband speech signals (16kHz sampling rate), when the frame length is set as 64 msec (1024 samples) and the frame shift is set as 16 msec (256 samples), respectively, we can design the bit allocation scheme as is shown in Tab.1. We can see that speech coding at different bit-rates could be flexibly implemented given the corresponding bit allocation schemes.

When the bit-rate is 1000 bit/s, there are totally 16 bits for each speech frame, so only 4 bits are allocated for scalar quantization of energy parameter z_0 and the last 12 bits are allocated for direct vector quantization of formant and pitch parameters $z_2 \sim z_{80}$. When the bit-rate is 2000 bit/s, there are totally 32 bits for each speech frame, so 6 bits are allocated for scalar quantization of energy parameter and another 26 bits are allocated for quantization other parameters.

In order to reduce the codebook searching complexity at the bit-rate of 2000 bit/s, we use multistage vector quantization (MSVQ) method to encode $z_2 \sim z_{80}$ efficiently. To make a tradeoff between the quantizing distortion and codebook searching burden, 2 cascaded codebooks are trained and the codebook at each stage consists of 2^{13} codewords, the quantization result of $z_2 \sim z_{80}$ is computed by comparison on quantization distortion of different combination of the reserved codewords at each stage.

3. EXPERIMENTS AND RESULTS

3.1. Dataset and Evaluation Metrics

We carry our experiments on the widely used TIMIT corpus to evaluate the performance of the proposed vocoder. TIMIT is a multi-speaker corpus, which contains 462 speakers in the training dataset and 168 speakers in the test dataset. At the training stage, the whole TIMIT training set with 4620 utterances were used for extracting mel-spectrograms and training the WaveGlow network model, the duration of the training speech is ∼4 hours. At the test stage, we chosen 300 utterances spoken by a total of 30 speakers from the test dataset for speech coding, the duration of test speech is ∼16 minutes. All the speech waveforms are sampled at 16kHz. The speech signal was enframed to 1024 samples using a hamming window and the frame shift is 256 samples. The dimension of MFCCs vector for each speech frame is 80, i.e., $K = 80$.

Two different objective metrics were used for evaluating the quality of coded speech. The first is perceptual evaluation of speech quality (PESQ) [17], which is adopted as the ITU-T P.862 standard and widely used for evaluating speech quality. Another is is the short-time objective intelligibility (STOI) [18], which is also a popular objective measure. PESQ demonstrates the overall speech quality while the STOI measure illustrates the speech intelligibility. For both the metrics, higher score indicates better performance. Also, we will take some subjective listening experiments to further demonstrate the performance of the proposed method.

3.2. Hyper-parameters Setting for WaveGlow Training

WaveGlow model was usually trained on single-speaker corpus for speech synthesis in previous study. However, speech coding for multi-speakers is much usual in real-word communication applications. Therefore, to obtain a good multi-speaker WaveGlow model on TIMIT corpus, the hyper-parameters should be carefully configured. Considering both the performance of WaveGlow network and the capacity of our hardware platform (Intel Xeon CPU (2.2GHz), 128G RAM and NVIDIA GeForce GTX 1080Ti $\times 2$ GPUs), we configured the hyper-parameters of WaveGlow as is shown in Tab.2. The quantized and unquantized mel-spectrogram were independently used as the input for WaveGlow training, the ADAM algorithm was chosen as the optimizer with the learning rate as 1×10^{-4}. After 1,110,000 epoches of training, we obtained a WaveGlow network model which was used as a decoder for low bit-rate speech coding.

3.3. Evaluation of speech quality

For simplicity, we denote the proposed speech coding algorithm via quantization of MFCCs and WaveGlow as WaveGlow codec, some other notations are as follows,

- OS: original speech signal
- UQ: speech synthesis from unquantized MFCCs
- UQT2000: WaveGlow codec at 2000 bit/s with unquantized

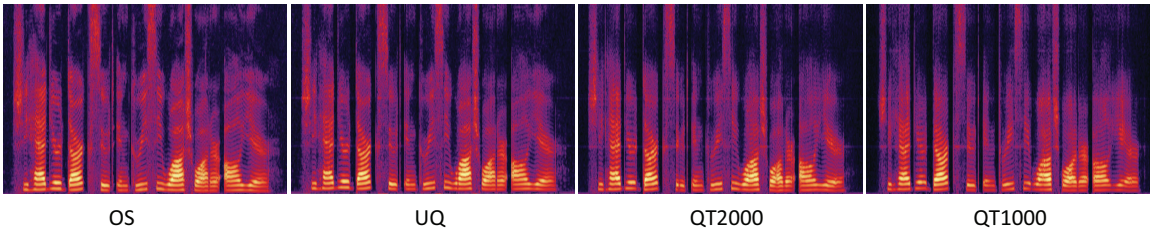

Fig. 2. Comparison on spectrograms of the TIMIT utterance "She had your dark suit and greasy wash water all year".

Table 2. hyper-parameters setting for WaveGlow training

hyper-parameter	value
number of flows	12
number of mel-channels	80
number of groups	8
number of layers for coupling module	8
number of mel-channels for coupling module	256
kernel size for coupling module	3
learning rate	1×10^{-4}
batch size	12

MFCCs as input for training WaveGlow model
• UQT1000: WaveGlow codec at 1000 bit/s with unquantized MFCCs as input for training WaveGlow model
• QT2000: WaveGlow codec at 2000 bit/s with quantized MFCCs as input for training WaveGlow model
• QT1000: WaveGlow codec at 1000 bit/s with quantized MFCCs as input for training WaveGlow model

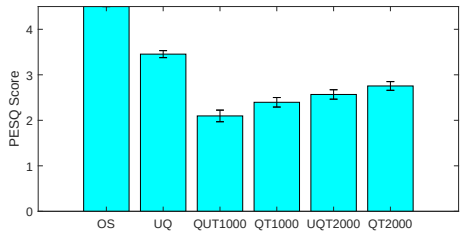

Fig. 3. speech quality in terms of PESQ score.

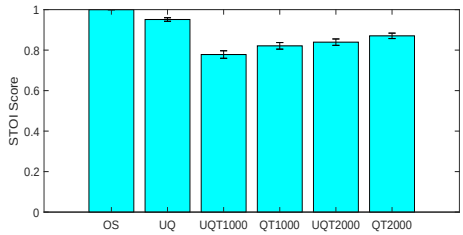

Fig. 4. speech quality in terms of STOI score.

Fig.2 shows the spectrograms of the reconstructed speech via WaveGlow codec for a typical TIMIT utterance. We can see that the structure of harmonic and frequency formant is both well preserved, which demonstrates that the original speech and the coded speech sounds closely. Fig.3 and Fig.4 shows the speech quality in terms of PESQ and STOI scores for the test set. It should be noted that WaveGlow trained with quantized MFCCs performs better than WaveGlow trained with unquantized MFCCs, because it overcomes the dismatch problem during the WaveGlow training and inference stage. We can also see that the output speech quality for QT2000 and QT1000 is acceptable as the PESQ scores of the output speech is about 2.75 and 2.52, respectively. We listened these coded speech signals and found that the output speech of WaveGlow codec preserves high intelligibility and somewhat naturalness though few audible artifacts exist.

We also conducted subjective listening tests. 10 volunteers rated the coded speech through the standard five point mean opinion score (MOS) [19]. Each volunteer was presented with 20 speech files encoded by WaveGlow codec and MELPe codec. The results are illustrated in Fig.5, which illustrates that WaveGlow codec provide substantially improved speech quality than MELPe codec at similar bit-rate. In detail, the MOS score for QT2000 and QT1000 is about 3.25 and 2.96, respectively.

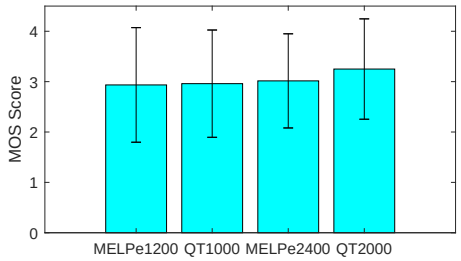

Fig. 5. speech quality in terms of MOS score.

4. CONCLUSIONS

This paper presented a new low bit-rate wideband speech coding approach though vector quantization of MFCCs. WaveGlow was used as a decoder in order to provide efficient and

high-quality speech coding at 1000-2000 bit/s. Experimental results demonstrate that WaveGlow codec is promising for low bit-rate source coding of speech signals with high speed inference. In further, other efficient generative models conditioned on mel-spectrogram, such as generative adversarial networks (GANs) [20][21], are also worth being explored for speech coding purpose. Moreover, the post-filtering technique is also worth studying to reduce the audible artifacts.

5. ACKNOWLEDGE

This work is partially supported by Natural Science Foundation of China(61701535, 61871471) and Key Research and Development Project of Shannxi Province (2020GY-015).

6. REFERENCES

[1] T. E. Tremain, "The government standard linear predictive coding algorithm: LPC10," *Speech Technol.*, vol. 1, pp. 40-49, 1982.

[2] M. R. Schroeder, B. S. Atal, "Code-excited linear prediction (CELP): High-quality speech at very low bit rates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1985, pp. 937–940.

[3] A. V. McCree and T. P. Barnwell,"Mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Trans. on Speech and Audio Process.*, vol. 3, no. 4, pp. 443–445, 1995.

[4] L. E. Boucheron, P. L. De Leon, and S. Sandoval, "Low bit-rate speech coding through quantization of mel-frequency cepstral coefficients," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 610-619, Feb. 2012.

[5] G. Min, X. W. Zhang, X. Zou, *et al*, "Perceptually weighted analysis-by-synthesis vector quantization for MFCC codec," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1379-1383, Oct. 2016.

[6] D. W. Griffin and J. S. Lim, "Signal estimation from modified short time fourier transform," *IEEE Trans. Acoustic, Speech, and Signal Process.*, vol. 32, no. 2, pp. 236-243, Apr. 1984.

[7] T. Bäckström,"End-to-end optimization of source models for speech and audio coding using a machine learning framework," in *Proc. Interspeech*, 2019, pp. 3401-3405.

[8] W. B. Kleijn, F. S. C. Lim, A. Luebs, *et al*, "Wavenet based low bit rate speech coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2018, pp. 676-680.

[9] C. Gârbacea, A. V. D. Oord, Y. Li, *et al*, "Low bit-rate speech coding with VQ-VAE and a WaveNet decoder," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2019, pp. 735-739.

[10] J. Klejsa, P. Hedelin, C. Zhou, *et al*, "High quality speech coding with simple RNN," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2019, pp. 7155-7159.

[11] J. Valin and J. Skoglund, "LPCNET: Improving Neural Speech synthesis through linear prediction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2019, pp. 5891-5895.

[12] R. Fejgin, J. Klejsa, L. Villemoes, *et al*, "Source coding of audio signals with a generative model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2020, pp. 341-345.

[13] G. Min, C. Q. Zhang, X. W. Zhang, *et al*, "Deep vocoder: low bit rate speech compression with deep autoencoder," in *Proc. IEEE Int. Conf. Multi. and Expo Workshops (ICMEW)*, 2019, pp. 1-6.

[14] H. Yalim Keles, J. Rozhon, H. Gokhan Ilk, *et al*, "DeepVoCoder: a CNN model for compression and coding of narrow band speech," in *IEEE ACCESS*, vol. 7, pp. 75081-75089, Jun. 2019.

[15] R. Prenger, R. Valle, and B. Catanzar, "Waveglow: a Flow-based generative network for speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2019, pp. 7-12.

[16] P. Govalkar, J. Fischer, F. Zalkow, *et al*, "A comparison of recent neural vocoders for speech signal reconstruction," in *Proc. 10th ISCA Speech Synthesis Workshop (SSW)*, 2019, pp. 7-12.

[17] A. W. Rix, J. G. Beerends, M. P. Hollier, *et al*, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2001, vol. II, pp. 749-752.

[18] C. H. Taal, R. C. Hendriks, R. Heusdens, *et al*, "An Algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Acoustic, Speech, and Signal Process.*, vol. 19, no. 7, pp. 2125-2136, Jul. 2011.

[19] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC, 2007.

[20] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al*, "Generative adversarial networks," *arXiv preprint*, arXiv:1406.2661, pp. 1-9, 2014.

[21] K. Kumar, R. Kumar, T. de Boissiere, *et al*, "MelGAN: generative adversarial networks for conditional waveform synthesis," in *Neural Information Processing Systems (NeurIPS)*, 2019, pp. 1-12.