ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746296

END-TO-END NEURAL SPEECH CODING FOR REAL-TIME COMMUNICATIONS

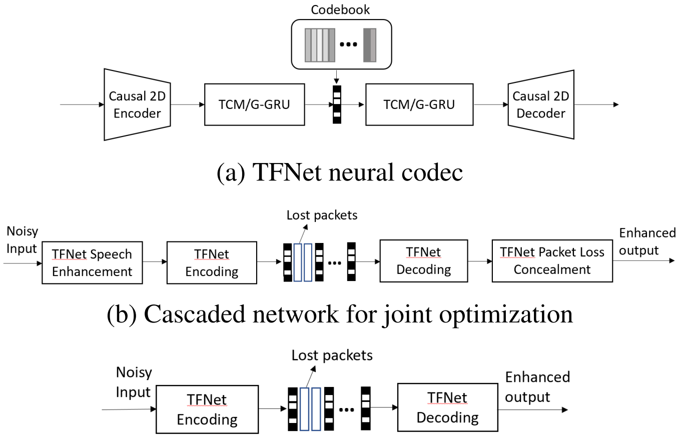
Xue Jiang[1*], *Xiulian Peng*[2], *Chengyu Zheng*[1], *Huaying Xue*[2], *Yuan Zhang*[1], *Yan Lu*[2]

[1] Communication University of China, Beijing, China
[2] Microsoft Research Asia, Beijing, China

ABSTRACT

Deep-learning based methods have shown their advantages in audio coding over traditional ones but limited attention has been paid on real-time communications (RTC). This paper proposes the TFNet, an end-to-end neural speech codec with low latency for RTC. It takes an encoder-temporal filtering-decoder paradigm that has seldom been investigated in audio coding. An interleaved structure is proposed for temporal filtering to capture both short-term and long-term temporal dependencies. Furthermore, with end-to-end optimization, the TFNet is jointly optimized with speech enhancement and packet loss concealment, yielding a one-for-all network for three tasks. Both subjective and objective results demonstrate the efficiency of the proposed TFNet.

Index Terms— neural audio coding, real-time communications, speech enhancement, packet loss concealment



(a) TFNet neural codec

(b) Cascaded network for joint optimization

(c) All-in-one network for joint optimization

Fig. 1. TFNet neural codec and its joint optimization for real-time communications.

1. INTRODUCTION

Deep learning has shown its great advantages in a wide range of applications including natural language processing, computer vision, etc. In recent years, such gains are observed in neural audio coding as well. Existing neural audio codecs could mainly be grouped into two categories, generative decoder models [1, 2, 3, 4, 5, 6] and end-to-end neural audio coding [7, 8, 9, 10]. The former leverages generative models such as WaveNet [1], its variants WaveGRU in Lyra [2], WaveRNN in LPCNet [6] and SampleRNN [3] at the decoder for low-rate speech coding; While for encoding, handcrafted features are extracted from a speech signal, which are typically by a standard parametric codec or a transformed log mel-spectra. These algorithms have clearly outperformed traditional codecs at a very low rate for speech coding. Further, some researchers introduce VQ-VAE [11] into compression and opens the door for end-to-end neural speech coding [7, 8]. Learned phone embeddings are encoded and a WaveNet is used at the decoder conditioning on the embedding to restore speech signal. These methods typically have a high complexity with auto-regressive networks. Without restricted by speech, some researchers propose to use autoencoders with vector quantization for end-to-end neural audio coding [9, 10]. A typical encoder is employed to extract features for encoding with vector quantization and a decoder is utilized for recovering the original audio waveform from quantized features. This category makes a versatile neural audio codec possible; however, its potential is yet far from full exploration. Among all these methods, they mostly still target at coding efficiency with little attention on low latency and error resilience for real-time communications.

In light of the success of the encoder-temporal filtering-decoder paradigm based on causal convolutions commonly used for regression tasks in real-time audio signal processing [12, 13], in this paper we propose the TFNet, a low-latency neural speech codec with end-to-end optimization for real-time communications, following this paradigm (see Fig.1(a)). We show that this paradigm works well for low-latency neural speech coding. Particularly, an interleaved structure with causal temporal convolution module (TCM) and group-wise gated recurrent unit (G-GRU) is proposed for temporal filtering for joint long-term and short-term correlation exploitation. Although taking speech as an example in this paper, the TFNet could be easily extended to other audio types such as music as well.

Furthermore, we exploit the optimization of codec under two types of degradation commonly existing in real-time

[*] This work was done when Xue Jiang was an intern at MSRA.

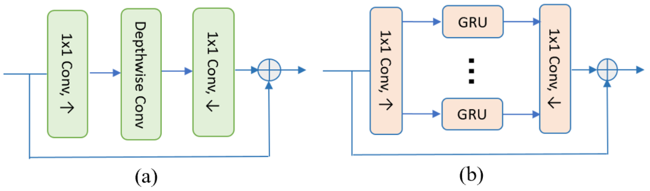


Fig. 2. Two causal temporal filtering modules. (a) Dilated temporal convolution module (TCM); (b) Group-wise gated recurrent unit (G-GRU).

communications, i.e. background noises and packet losses. We show that with the end-to-end optimization brought by audio coding, user-perceived audio quality could be optimized more efficiently by joint optimization instead of three separate TFNet models optimized for each local target. Particularly, we show that a single TFNet model performs comparatively with three cascaded models under joint optimization (see Fig.1). This investigation also shows the efficiency of TFNet for both separation (speech enhancement) and restoration (packet loss concealment and coding) tasks.

The rest of the paper is organized as follows. Section 2 expains the details of the proposed TFNet codec. The joint optimization with speech enhancement and packet loss concealment is described in Section 3. Section 4 shows the experimental results and finally Section 5 concludes this paper.

2. PROPOSED TFNET CODEC

2.1. Overview

The TFNet-based codec takes the time-frequency spectrum (20ms window with a 5ms hop length) as input with a power-law compression on the amplitude before feeding into the network. As the dynamic range of speech is high due to harmonics, the compression performs as a kind of input normalization so that the importances of different frequencies are balanced and the training is more stable.

The network consists of an encoder and several temporal filtering blocks for encoding (see Fig.1(a)). Among them, the encoder exploits local two-dimensional (2D) correlations and the temporal filtering blocks exploit longer-term temporal dependencies with past frames for feature extraction. The two-level feature extraction is important in learning to extract features with good representation capability, error resilience to packet losses and to possibly remove undesired information such as background noises if wanted. The learned features are then quantized through a learned vector quantizer and coded in fixed-length coding.

For decoding, there are several temporal filtering blocks followed by a decoder for reconstruction. Considering the packet losses in real-time communications, the decoding needs to be resilient to this losses with recovery capability and minimum error propagation. Therefore, a heterogeneous structure is designed with more temporal filtering blocks for decoding than encoding.

The whole network is end-to-end trained to optimize the reconstruction quality under a rate constraint. All convolutions are causal in temporal dimension so that it can keep a low latency of 20ms.

2.2. Encoder and Decoder

The encoder consists of several causal 2D convolutional layers, each followed by a batch normalization (BN) and a parametric ReLU (PRelu) for nonlinearity. After each convolutional layer, the feature is downsampled by 2 or 4 in frequency dimension and finally all frequency information is folded into channels. Let $X^I \in R^{T\times F\times 2}$ denote the input feature. After the encoder, the feature is $X^E \in R^{T\times 1\times C}$ for feeding into temporal filtering blocks. T, F and C are number of frames, frequency bins and channels, respectively. Convolutions are causal along the temporal dimension so T is kept without any downsampling. The decoder is symmetric to the encoder with causal 2D deconvolutional layers. After the decoder, it outputs the reconstructed complex spectrum $X^R \in R^{T\times F\times 2}$ with an inverse STFT to get the output waveform.

2.3. Temporal Filtering

We utilize two types of temporal filtering blocks as shown in Fig.2, the dilated temporal convolution module (TCM) and the group-wise gated recurrent unit (G-GRU). Both are causal and low-complexity. The TCM module as that in [14] consists of two convolutions with a kernel size of 1×1 to change channel dimensions and dilated depthwise convolutions to exploit temporal correlations with low complexity. Several TCM blocks with different dilation rates are grouped as a large block to increase the receptive field and diversities. The group-wise GRU blocks in Fig.2 (b) split channels into N groups and leverage temporal dependencies inside each group independently. This group-wise GRU variants not only reduce the complexity but also increase the flexibility and representation capability for providing frequency-aware temporal filtering as channels are learned from frequencies.

As TCM is more efficient in exploring short-term and middle-term temporal evolutions while GRU is mostly proposed to capture long-term dependencies, we propose to combine them in an interleaved way for capturing short-term and long-term temporal correlations at different depths. Our experimental results verify that the interleaved structures are more efficient than a single one, which will be presented in Section 4.

2.4. Vector Quantization

The vector quantizer discretizes the learned features in encoding with a set of learnable codebooks according to the target bitrate. Before quantization, the features after encoding $X^S \in R^{T\times 1\times C}$ are reduced to $X^Q \in R^{T\times 1\times C'}$ through a

1×1 convolution ($C' < C$). We take a group quantization by splitting channels C' into N groups and coding each group by an independent codebook. Let S denote the number of codewords in each codebook and $K = C'/N$ the dimension of each codeword. In the proposed scheme, a window length of 20ms and hop length of 5ms is adopted for STFT and thus the bitrate is given by $N \times log_2 S/5$ kbps. For 6kbps, C', N, S and K are set to 120, 3, 1024, and 40, respectively. The codebooks are learned with exponential moving average, following that in [11]. The quantized features $\hat{X}^Q \in R^{T \times 1 \times C'}$ are enlarged to the shape $T \times 1 \times C$ before feeding into the temporal filtering blocks in decoding.

2.5. Loss Function

The loss function is a combination of two terms $\mathcal{L} = \mathcal{L}_{recon} + \alpha \mathcal{L}_{VQ}$. $\mathcal{L}_{recon}$ is the reconstruction loss and the other $\mathcal{L}_{VQ}$ puts a constraint on vector quantization. We use a mean-square error on the power-law compressed spectrum between the original and the decoded signals for reconstruction loss [15]. To ensure STFT consistency [16], the decoded spectrum is first transformed into waveform domain through an inverse STFT and then transformed into TF-domain again through a STFT to calculate the loss. The second term $\mathcal{L}_{VQ}$ is the commitment loss used in VQ-VAE, which forces the encoder to generate a representation close to its codeword. α is a weighting factor to balance the two terms.

3. JOINT OPTIMIZATION WITH SPEECH ENHANCEMENT AND PACKET LOSS CONCEALMENT

In real-time communications, there are several types of degradations besides quality loss by audio coding, such as background noises and packet losses. Owing to the end-to-end learnable audio codec, it is feasible to jointly optimize the audio coding with speech enhancement (SE) and packet loss concealment (PLC). We consider two ways of joint optimization, the cascaded network with an enhancer before the codec and a PLC network after it (see Fig.1(b)), and the all-in-one network that takes the similar network structure as the codec but optimized for noisy input with packet losses (see Fig.1(c)).

3.1. Cascaded Network

The cascaded network consists of three modules, an enhancer for pre-processing, an audio codec and a PLC network for post-processing. As speech is more efficient in compression than a noisy audio, the enhancer is put before the codec. The three modules are all based on TFNet-like structures and jointly trained in an end-to-end way.

The pre-processing enhancer takes noisy audio as input and outputs enhanced audio for feeding into the codec. Different from the TFNet-based codec, there are skip connections between the encoder and the decoder in the enhancer to get rid of information loss. We employ causal gated blocks in decoder and output an amplitude gain and the phase for reconstruction, similar to that in [17].

Under packet losses, the neural codec is adjusted in that in decoding it takes both the quantized features with lost packets as zero and a mask showing where the loss happens as input. The mask is also injected into each temporal filtering blocks in decoding. The post-processing PLC module operates in the waveform domain, taking a TFNet-based structure with both the decoded audio and the mask as input. There are also skip connections in the PLC network as that in enhancer. As a restoration task, it outputs a complex residue in TF domain which is added into the spectrum of the decoded audio for reconstruction.

For training, the three networks are concatenated and jointly trained from end to end. For better quality, we take a two-stage training. First, the enhancer and the codec are separately trained with noisy and clean data, respectively. Then the cascaded network is finetuned from that, with two additional supervisions at the output of the enhancer and the codec, respectively, using the same reconstruction loss as $\mathcal{L}_{recon}$.

3.2. All-in-One Network

The all-in-one network is resilient to both background noises and packet losses with only a single codec network. To accommodate packet losses, the decoding part in the codec is adjusted similar to that in the cascaded network. It is trained from scratch with an auxiliary supervision added for the encoding part to remove noises for efficient coding. This is achieved by adding a decoder after the temporal filtering blocks of the encoding, which is forced to output clean audio in training. During inference, this decoder is not needed.

4. EXPERIMENTAL RESULTS

4.1. Datasets, Settings and Evaluation Metrics

We synthesized 890 hours of 16khz noisy audios with clean speech, noises and room impulses from the Deep Noise Suppression Challenge at ICASSP 2021 [18]. The clean audio includes multilingual speech, emotional and singing clips. The SNR is randomly chosen inbetween -5dB and 20dB and the speech level within -40 to -10dB. Each audio is cut into 3-second segments for training. The speech enhancement performs both denoising and dereverberation. The packet losses are simulated following the three-state model [19]. For testing, we take 1400 audios each with 10 seconds long without any overlapping with training data.

During training, we use the adam optimizer with a learning rate of 0.0004. The network is trained for 100 epochs with a batch size of 200.

In evaluation, except the subjective listening test in Section 4.2, three metrics are used for ablation studies in Sec-

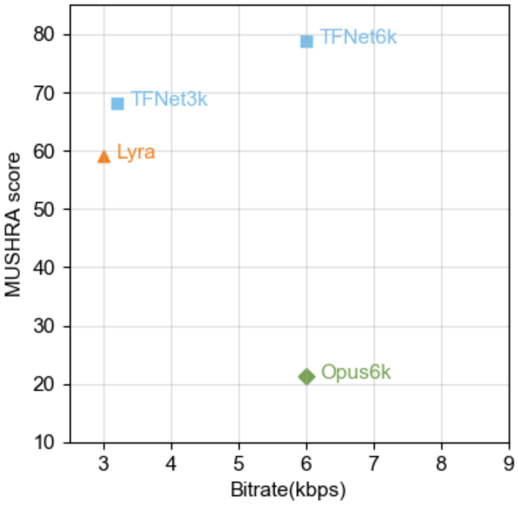


Fig. 3. Subjective evaluation results (Reference score: 88).

tion 4.3 and 4.4, the wideband PESQ [20], STOI [21] and the DNSMOS [22]. Although they are not designed and optimized for exactly the same task, we found that for the same kind of distortions in all compared schemes, they match well with perceptual quality.

4.2. Comparison with other Codecs

The codec network is trained and measured on the clean data only in this section. We conduct a subjective listening test with a MUSHRA-inspired crowd-sourced method [23]. There are 10 participants, each evaluates 12 samples. We compare the TFNet-based neural codec with Lyra* and Opus†, two codecs used for real-time communications. The former is a neural speech codec proposed in this year. As shown in Fig.3, the TFNet at around 3kbps clearly outperforms Lyra at 3kbps and TFNet at 6kbps is much better than Opus at 6kbps, which demonstrates its superiority.

4.3. Evaluation on Joint Optimizations

We evaluate the joint optimization of codec, speech enhancement and PLC in this section using the noisy/clean paired data with simulated packet loss traces. Three methods are compared here, a baseline with separately trained enhancement, coding and PLC models, the cascaded network and the all-in-one network. In baseline, coding and PLC networks are trained on raw clean data only. The enhancer and PLC networks have 470K parameters and 1.2M MACs per 20ms, far less than the codec network with 5M parameters.

In Table 1 and 2, the comparison results on two and three-task joint optimizations are presented, respectively. It is observed that the two joint optimization methods clearly outperforms the baseline in all metrics. Although no pre-processing nor post-processing networks are used, the all-in-one network performs competitively with the cascaded one, showing the strong discrimination and representation capability of TFNet. Another observation is that the PLC network trained on raw clean data in baseline method is sensitive to mismatch in the input.

Table 1. Comparison of coding efficiency under background noises.

	PESQ	STOI	DNSMOS
Baseline	1.740	0.811	3.29
Cascaded	1.777	**0.822**	**3.51**
All-in-one	**1.794**	0.821	3.48

Table 2. Comparison of coding efficiency under background noises and packet losses.

	PESQ	STOI	DNSMOS
Baseline	1.413	0.747	3.04
Cascaded	**1.545**	**0.778**	**3.38**
All-in-one	1.510	0.770	3.33

4.4. Evaluation on Temporal Filtering Modules

We compare the interleaved structure in TFNet neural codec with two modules TCM and GRU commonly used in regression tasks of speech enhancement here. All schemes are compared under the same computational complexity with 1.4M parameters and 3.3M MACs for each 20ms window for encoding and decoding. In this study, all temporal filtering modules are used for decoding only to evaluate their recovery capability. Table 3 shows the comparison results. It can be seen that the interleaved structure performs the best for capturing both short-term and long-term temporal correlations.

Table 3. Comparison of different temporal filtering modules.

	PESQ	STOI
TCM	2.447	0.869
GRU	2.360	0.863
Interleave-TCM-G-GRU	**2.501**	**0.870**

5. CONCLUSIONS

We propose the TFNet, a low-latency neural speech codec with end-to-end optimization. Taking it as a backbone, we further investigate the joint optimization of the codec with speech enhancement and packet loss concealment. Experimental results demonstrate the representation capability of TFNet in speech coding and regression tasks. As the network itself makes no assumption on the nature of the audio signal it encodes and no explicit speech-specific features are extracted, it can be extended for other audio types as well in the future.

*https://github.com/google/lyra
†https://opus-codec.org

6. REFERENCES

[1] W.B. Kleijin, F.S. Lim, A. Luebs, and J. Skoglund, "WaveNet based low rate speech coding," in *ICASSP*. IEEE, 2018, pp. 676–680.

[2] W.B. Kleijn, A. Storus, M. Chinen, T. Denton, F.S.C. Lim, A. Luebs, J. Skoglund, and H. Yeh, "Generative speech coding with predictive variance regularization," in *arXiv:2102.09660*, 2021.

[3] J. Klejsa, P. Hedelin, C. Zhou, R. Fejgin, and L. Villemoes, "High-quality speech coding with sample RNN," in *ICASSP*. IEEE, 2019, pp. 7155–7159.

[4] R. Fejgin, J. Klejsa, L. Villemoes, and C. Zhou, "Source coding of audio signals with a generative model," in *ICASSP*. IEEE, 2020, pp. 341–345.

[5] J. Skoglund and J.M. Valin, "Improving Opus low bit rate quality with neural speech synthesis," in *Interspeech*, 2020.

[6] Valin J.M. and Skoglund J., "LPCNet: improving neural speech synthesis through linear prediction," in *ICASSP*. IEEE, 2019.

[7] C. Gârbacea, A. van den Oord, Y. Li, F.S. Lim, A. Luebs, O. Vinyals, and T. C. Walters, "Low bit-rate speech coding with VQ-VAE and a WaveNet decoder," in *2019 IEEE Int. Conf. Acoust Speech Signal Processing (ICASSP)*. IEEE, 2019, pp. 735–739.

[8] J. Williams, Yi Zhao, E. Cooper, and J. Yamagishi, "Learning disentangled phone and speaker representations in a semi-supervised VQ-VAE paradigm," in *2021 IEEE Int. Conf. Acoust Speech Signal Processing (ICASSP)*. IEEE, 2021.

[9] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "SoundStream: an end-to-end neural audio codec," in *arXiv:2107.03312v1*, 2021.

[10] K. Zhen, J. Sung, M.S. Lee, S. Beack, and M. Kim, "Cascaded cross-module residual learning towards lightweight end-to-end speech coding," in *Proceedings of the Annual Conference of the International Speech and Communication Association (Interspeech)*, 2019.

[11] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu, "Neural discrete representation learning," *arXiv preprint arXiv:1711.00937*, 2017.

[12] A. Pandey and D.L. Wang, "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *2019 IEEE Int. Conf. Acoust Speech Signal Processing (ICASSP)*. IEEE, 2019.

[13] K. Tan and D.L. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Proceedings of the Annual Conference of the International Speech and Communication Association (Interspeech)*, 2018.

[14] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[15] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *arXiv preprint arXiv:1804.03619*, 2018.

[16] Scott Wisdom, John R Hershey, Kevin Wilson, Jeremy Thorpe, Michael Chinen, Brian Patton, and Rif A Saurous, "Differentiable consistency constraints for improved deep speech enhancement," in *ICASSP*. IEEE, 2019, pp. 900–904.

[17] C. Zheng, X. Peng, Y. Zhang, S. Srinivasan, and Y. Lu, "Interactive speech and noise modeling for speech enhancement," in *AAAI*, 2021.

[18] C. K A Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "ICASSP 2021 deep noise suppression challenge," *arXiv preprint arXiv:2009.06122*, 2020.

[19] B.P. Milner and A.B. James, "An analysis of packet loss models for distributed speech recognition," *Proceedings INTERSPEECH, 8th Internaltional Conference on Spoken Language Processing*, 2004.

[20] ITUT Rec, "P.862.2: Wideband extension to recommendation p.862 for the assessment of wideband telephone networks and speech codecs," *International Telecommunication Union, CH–Geneva*, 2005.

[21] C.H.Taal, R.C.Hendriks, R.Heusdens, and J.Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *ICASSP*, 2010.

[22] C. K A Reddy, V. Gopal, and R. Cutler, "DNSMOS: a non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *2021 IEEE Int. Conf. Acoust Speech Signal Processing (ICASSP)*. IEEE, 2021.

[23] ITU-R, "Recommendation BS.1534-1: Method for the subjective assessment of intermediate quality level of audio systems," *International Telecommunication Union Radiocommunication Assembly*, 2001.