Scalable and Efficient Neural Speech Coding: A Hybrid Design

Kai Zhen, *Student Member, IEEE*, Jongmo Sung, Mi Suk Lee, Seungkwon Beack, and Minje Kim, *Senior Member, IEEE*

***Abstract*—We present a scalable and efficient neural waveform coding system for speech compression. We formulate the speech coding problem as an autoencoding task, where a convolutional neural network (CNN) performs encoding and decoding as a neural waveform codec (NWC) during its feedforward routine. The proposed NWC also defines quantization and entropy coding as a trainable module, so the coding artifacts and bitrate control are handled during the optimization process. We achieve efficiency by introducing compact model components to NWC, such as gated residual networks and depthwise separable convolution. Furthermore, the proposed models are with a scalable architecture, cross-module residual learning (CMRL), to cover a wide range of bitrates. To this end, we employ the residual coding concept to concatenate multiple NWC autoencoding modules, where each NWC module performs residual coding to restore any reconstruction loss that its preceding modules have created. CMRL can scale down to cover lower bitrates as well, for which it employs linear predictive coding (LPC) module as its first autoencoder. The hybrid design integrates LPC and NWC by redefining LPC's quantization as a differentiable process, making the system training an end-to-end manner. The decoder of proposed system is with either one NWC (0.12 million parameters) in low to medium bitrate ranges (12 to 20 kbps) or two NWCs in the high bitrate (32 kbps). Although the decoding complexity is not yet as low as that of conventional speech codecs, it is significantly reduced from that of other neural speech coders, such as a WaveNet-based vocoder. For wide-band speech coding quality, our system yields comparable or superior performance to AMR-WB and Opus on TIMIT test utterances at low and medium bitrates. The proposed system can scale up to higher bitrates to achieve near transparent performance.**

***Index Terms*—Neural speech coding, waveform coding, representation learning, model complexity.**

Manuscript received March 26, 2021; revised June 28, 2021 and September 9, 2021; accepted November 8, 2021. Date of publication November 19, 2021; date of current version December 20, 2021. This work was supported by the Institute for Information and Communications Technology Promotion (IITP) funded by the Korea government (MSIT) under Grant 2017-0-00072 (Development of Audio/Video Coding, and Light Field Media Fundamental Technologies for Ultra Realistic Tera-Media). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Tom Backstrom. *(Corresponding author: Minje Kim.)*

Kai Zhen is with the Department of Computer Science and Cognitive Science Program, Indiana University, Bloomington, IN 47408 USA (e-mail: zhenk@iu.edu).

Jongmo Sung, Mi Suk Lee, and Seungkwon Beack are with the Electronics and Telecommunications Research Institute, Daejeon 34129, Korea (e-mail: jmseong@etri.re.kr; lms@etri.re.kr; skbeack@etri.re.kr).

Minje Kim is with the Department of Intelligent Systems Engineering, Indiana University, Bloomington, IN 47408 USA (e-mail: minje@indiana.edu).

Digital Object Identifier 10.1109/TASLP.2021.3129353

I. Introduction

SPEECH coding can be implemented as an encoder-decoder system, whose goal is to compress input speech signals into the compact bitstream (encoder) and then to reconstruct the original speech from the code with the least possible quality degradation. Speech coding facilitates telecommunication and saves data storage among many other applications. There is a typical trade-off a speech codec must handle: the more the system reduces the amount of bits per second (bitrate), the worse the perceptual similarity between the original and recovered signals is likely to be perceived. In addition, the speech coding systems are often required to maintain an affordable computational complexity when the hardware resource is at a premium. For decades, speech coding has been intensively studied yielding various standardized codecs that can be categorized into two types: the vocoders and waveform codecs. A vocoder, also referred to as parametric speech coding, distills a set of physiologically salient features, such as the spectral envelope (equivalent to vocal tract responses including the contribution from mouth shape, tongue position and nasal cavity), fundamental frequencies, and gain (voicing level), from which the decoder *synthesizes* the speech. Typically, a vocoder operates at 3 kbps or below with high computational efficiency, but the synthesized speech quality is usually limited and does not scale up to higher bitrates [1]–[3]. On the other hand, a waveform codec aims to accurately reconstruct the input speech signal, which features up-to-transparent quality in a high bitrate range [4]. AMR-WB [5], for instance, can be seen as a hybrid waveform codec, because it employs speech modeling as in many other waveform codecs [6]–[8]. EVS [9], a recently standardized 3GPP voice and audio codec, has noticeably optimized frame error robustness, yielding a much-enhanced frame error concealment performance against than AMR-WB [10]. Similar to EVS, Opus, a waveform codec at its core, can also be applied to both speech and audio signals where it uses the LPC-based SILK algorithm for the speech-oriented model [11] and scales up to 510 kbps for transparent audio streaming and archiving.

Under the notion of unsupervised speech representation learning, deep neural network (DNN)-based codecs have revitalized the speech coding problem and provided different perspectives [12], [13]. The major motivation of employing neural networks to speech coding is twofold: to fill the performance gap between vocoders and waveform codecs towards a near-transparent speech synthesis quality; to use its trainable encoder

and learn latent representations which may benefit other DNN-implemented downstream applications, such as speech enhancement [14], [15], speaker identification [16] and automatic speech recognition [17], [18]. Having that, a neural codec can serve as a trainable acoustic unit integrated in future digital signal processing engines [13].

Recently proposed neural speech codecs have achieved high coding gain and reasonable quality by employing deep autoregressive models. The superior speech synthesis performance achieved in WaveNet-based models [19] has successfully transferred to neural speech coding systems, such as in [20], where WaveNet serves as a decoder synthesizing wideband speech samples from a conventional non-trainable encoder at 2.4 kbps. Although its reconstruction quality is comparable to waveform codecs at higher bitrates, the computational cost is significant due to the model size of over 20 million parameters.

Meanwhile, VQ-VAE [12] integrates a trainable vector quantization scheme into the variational autoencoder (VAE) [21] for discrete speech representation learning. While the bitrate can be lowered by reducing the sampling rate 64 times, the downside for VQ-VAE is that the prosody can be significantly altered. Although [22] provides a scheme to pass the pitch and timing information to the decoder as a remedy, it does not generalize to non-speech signals. More importantly, VQ-VAE as a vocoder does not address the complexity issue since it uses WaveNet as the decoder. Although these neural speech synthesis systems noticeably improve the speech quality at low bitrates, they are not feasible for real-time speech coding on the hardware with limited memory and bandwidth.

LPCNet [23] focuses on efficient neural speech coding via a WaveRNN [24] decoder by leveraging the traditional linear predictive coding (LPC) techniques. The input of the LPCNet is formed by 20 parameters (18 Bark scaled cepstral coefficients and 2 additional parameters for the pitch information) for every 10 ms frame. All these parameters are extracted from the non-trainable encoder, and vector-quantized with a fixed codebook. As discussed previously, since LPCNet functions as a vocoder, the decoded speech quality is not considered transparent [1].

In this paper, we propose a novel neural speech coding system, with a lightweight design and scalable performance. First, we design a generic neural waveform codec with only 0.35 million parameters where 0.12 million parameters belong to the decoder. Compared to our previous models in [25], [26] where the decoder has 0.23 million parameters, the current neural codec employs gated linear units to boost the gradient flow during model training and depthwise separable convolution to achieve further efficiency during decoding, as detailed in Section II. Based on this neural codec, our full system features two mechanisms to integrate speech production theory and residual coding techniques in Section III. Benefited from the residual-excited linear prediction (RELP) [27], we conduct LPC and apply the neural waveform codec to the excitation signal, which is illustrated in Section III-A. In this integration, a trainable quantizer bridges the encoding of linear spectral pairs and the corresponding LPC residual, making the speech coding pipeline end-to-end trainable. We also enable residual coding among neural waveform codecs to scale up the performance for high bitrates (Section III-B). In summary, the proposed system has following characteristics:

- *Scalability:* Similar to LPCNet [23], the proposed system is compatible with conventional spectral envelope estimation techniques. However, ours operates at a much wider bitrate range with comparable or superior speech quality to standardized waveform codecs.
- *Compactness:* The neural waveform codec in our system is with a much lower complexity than WaveNet [19] and VQ-VAE [12] based codecs. Our decoder contains only 0.12 million parameters which is $160\times$ more compact than a WaveNet counterpart. Our TensorFlow implementation's execution time to encode and decode a signal is only 42.44% of its duration on a single-core CPU in the low-to-medium bitrates and 80.21% in the high bitrate, facilitating real-time communications.
- *Trainability:* Our method is with a trainable encoder as in VQ-VAE, which can be integrated into other DNNs for acoustic signal processing. Besides, it is not constrained to speech, and can be generalized to audio coding with minimal effort as shown in [28].

TABLE I
CATEGORICAL SUMMARY OF RECENTLY PROPOSED NEURAL SPEECH CODING SYSTEMS. ✔ MEANS THE SYSTEM SUPPORTS THE FEATURE WHILE ✗ DOES NOT. ● MEANS IT IS NOT KNOWN

	WaveNet [20]	VQ-VAE [22]	LPCNet [23]	Proposed
Transparent coding	✔	●	✗	✔
Less than 1M parameters	✗	✗	✔	✔
Real-time communications	✗	✗	✔	✔
Encoder trainable	✔	✔	✗	✔

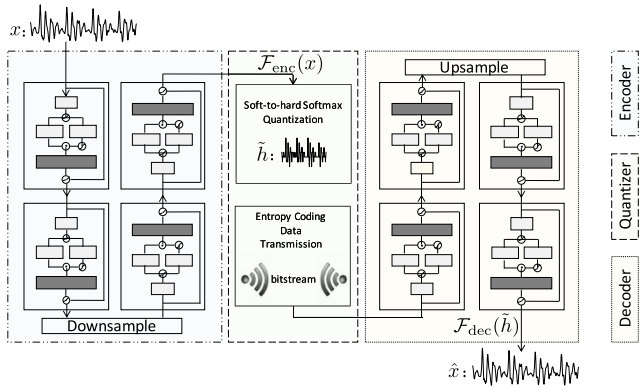
Table I highlights the comparison to the other existing neural speech codecs.

This paper is an extension of the authors' previous conference presentations [25], [26], where some initial ideas were already discussed. The new contributions presented this journal version are listed as follows:

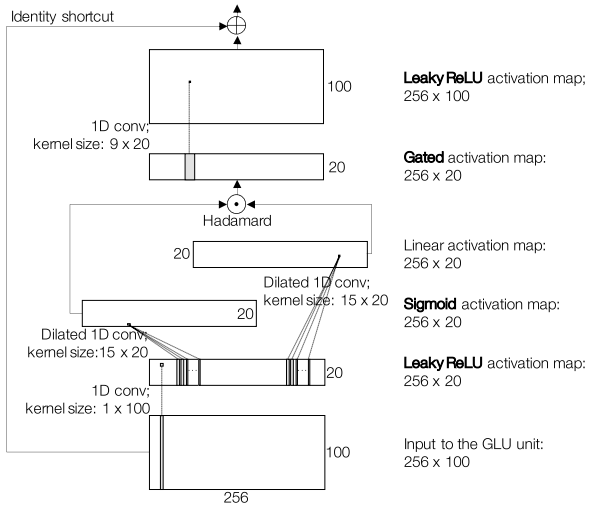
- *Novel Algorithmic Enhancements:* We propose new neural network architectures to form a new baseline autoencoder module and used it everywhere in our codecs. In our previous works, we have used a 1D convolutional neural network (CNN) that defines an autoencoder block with an identity shortcut as in the ResNet architecture [29]. While this architecture has been effective, in this journal paper, we propose to use the dilated gated linear units and depthwise separable convolution to reduce the kernel size without inducing any performance degradation. Consequently, our NWC is defined by 0.35 M parameters, whose decoder part accounts for only 0.12 M parameters. Compared to our previous models that are already small with only 0.45 M parameters, the newly introduced reduction amounts to 22.2%. If we only compare the decoder parts, it is a 47.8% reduction. Although the proposed architecture is more compact than our previous models or the WaveNet-based codecs, since neural codecs' complexity is much larger than the traditional speech codecs, the additional model

complexity reduction with no degradation of performance is promising. The architectural improvement are presented in Section II-A.

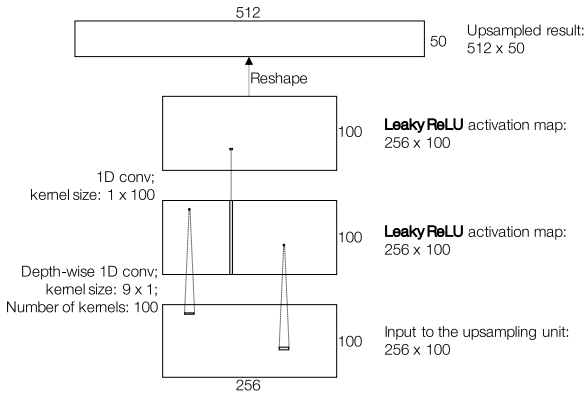
- *Extensive Experimental Validation:* In our previous works, the experimental validation was to prove the initial concepts individually proposed each paper. In this time, we conduct an extensive and thorough experiments to provide the readers with a full view to the whole building-blocks of the neural speech coding. To this end, we define four candidate systems, from Model-I to IV, by incrementally adding new modules, such as LPC, the trainable LPC quantizer, and multiple concatenated neural autoencoders. The objective and subjective tests validate each of these additions in a full view (Table III and Fig. 7).
- *Additional Analyses and Ablation Tests:* We also provide detailed experimental validation for most of the claims made in the paper by designing and performing separate experiments, which were missing in the previous papers.
  - Section IV-D1 provides experimental verification that the proposed compact neural architecture does not induce performance loss.
  - Section IV-D2 presents a detailed analysis of the behavior of the cascaded autoendoers and the impact of different training strategies.
  - Section IV-F1 explores contribution of different loss terms in our training objective by performing ablation tests, and then proposes an optimal combination of hyperparameters.
  - Section IV-F2 also conducts an ablation test to empirically verify that the proposed trainable LPC quantization algorithm improves speech quality at the same bitrate.
  - Section IV-F3 and IV-F4 analyze the bit allocation behavior among the different submodules. Since the bit allocation strategy is decided by the learning algorithm, these analyses provide evidence that our models dynamically adapt to the characteristics of the signals given the limited bit budget.
  - Section IV-G presents additional analyses on computational complexity and execution time ratios to discuss the potential of the neural codecs in real-time applications.
  - Last but not least, in Section IV-G3 we discuss the implementation issues and the limitations of the proposed system in the context of real-world application scenarios.



(a) The high-level structure of proposed neural waveform codec



(b) Dilated gated linear unit (GLU)



(c) Depthwise separable 1D convolution for upsampling

Fig. 1. The proposed architecture for lightweight NWC.

II. End-to-End Neural Waveform Codec (NWC)

The neural waveform codec (NWC), is an end-to-end autoencoder that forms the base of our proposed coding systems. NWC directly encodes the input waveform $\boldsymbol{x} \in \mathbb{R}^T$ using a convolutional neural network (CNN) encoder $\mathcal{F}_{\text{enc}}(\cdot)$, i.e., $\boldsymbol{h} \leftarrow \mathcal{F}_{\text{enc}}(\boldsymbol{x})$. Then, the quantization process $\mathcal{Q}(\cdot)$ converts the encoding into a bitstring $\tilde{\boldsymbol{h}} \in \mathbb{R}^N$, which is followed by lossless data compression and bitstream transmission. On the receiver side, the decoder reconstructs the waveform as $\boldsymbol{x} \approx \hat{\boldsymbol{x}} \leftarrow \mathcal{F}_{\text{dec}}(\tilde{\boldsymbol{h}})$. Fig. 1(a) depicts NWC's overall system architecture. The structure is detailed in Table II. It serves as a basic component in the proposed speech coding system in Section III. In this section, we first introduce the architectural improvement that reduced our model's complexity. Next, we also introduce two strategies