INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Multi-mode Neural Speech Coding Based on Deep Generative Networks

Wei Xiao[1], *Wenzhe Liu*[1], *Meng Wang*[1], *Shan Yang*[2], *Yupeng Shi*[1], *Yuyong Kang*[1], *Dan Su*[2], *Shidong Shang*[1], *Dong Yu*[3]

[1]Tencent Ethereal Audio Lab, Tencent, Shenzhen, China
[2]Tencent AI Lab, Tencent, Shenzhen, China
[3]Tencent AI Lab, Tencent, Bellevue, WA, USA

{denniswxiao, wenzheliu, markuswang, shaanyang, yupengshi, yuyongkang, dansu, simeonshang, dyu}@tencent.com

Abstract

The wideband or super wideband speech is one of the most prominent features in real-time communication services, with higher resolution spectrum. However, it requires higher computing expenses. In this paper, we introduce the Penguins codec, based on a multi-mode neural speech coding structure that combines sub-band speech processing and applies different strategies from the low band to the high band. Especially, it refers to deep generative networks with perceptual constraint loss functions and knowledge distillations to reconstruct wideband components and bandwidth extension to generate artificial super wideband components. The method results in high-quality speech at very low bitrates. Several subjective and objective experiments, including ablation studies, were organized, and the results proved the merit of the proposed scheme when compared with traditional coding schemes and state-of-the-art neural coding methods.

Index Terms: speech coding, quadrature mirror filter, deep generative model, bandwidth extension

1. Introduction

Speech coding is one of the fundamental technologies in voice telephony and real-time communication (RTC) services. It refers to speech feature estimation using audio signal processing and feature quantization by general data compression methods at the sending side; the received features are extracted and used to synthesize the speech at the receiving side. It improves the customers' experience in miscellaneous RTC applications (mobile telephony, social chatting, etc.) and boosts technology improvement in the meantime.

The PCM digital telephony codec, e.g., ITU-T G.711 [1], is a narrow band waveform-based technique that adopts logarithmic companding laws to express each sample by 8 bits, and this serie extends to support 16 kHz sampling rate later [2].

The linear prediction coding (LPC) is another approach widely applied in speech coding, including coded-excited linear prediction (CELP) [3]. The CELP is composed of a linear prediction stage that models the spectral envelope of the speech and a codebook-based excitation model that expresses the residual of the LPC. At the encoder, the CELP performs a search procedure with perceptually weighted constraints to obtain the best parameters, and related index of codebooks. At the decoder, both excitation and LP coefficients are obtained from the received index of codebooks; then, the speech is reconstructed by LP synthesis filtering. The CELP has been widely adopted by different international standards [4-6].

Compared to PCM and LPC approaches, the transform coding is another branch of speech coding that always transforms the speech into the frequency domain; then, the frequency coefficients are divided into multiple bands. For each band, the average power is calculated as the spectrum envelope. The spectrum envelopes and the normalized coefficients are quantized and compressed by entropy coding. At the decoder, the received spectrum envelope and normalized coefficients are combined to generate frequency coefficients, and then the inversed transform is applied to reconstruct the speech. It is noted that some psychoacoustic principles are used to improve coding efficiency [7].

Further, the super wideband (SWB) speech coding was also investigated due to the business requirements for higher resolution of the speech in RTC services. Such SWB methods combine LPC, transform coding, sub-band coding [8], and bandwidth extension [9-10] to realize high-efficiency coding of SWB speech signals.

The foundation of the above methods includes generalized speech analysis, processing, and synthesis (SP-based). In recent years, the data-driven method has also become a research focus in speech coding. The involvement of so-called deep neural networks (neural-based) in speech coding can bring out better quality than SP-based approaches when the bitrate is less than 8kbps.

The advance of the generative model realizes waveform or parametric coding by using neural synthesis [11]. Later, the LPCNet demonstrates how to combine linear prediction and recurrent neural networks to synthesize the speech with moderate computational consumption [12]. Recently, the Lyra feeds quantized log Mel spectra to WaveGRU and outperforms SP-based methods at low rates [13]. Another neural-based approach, the so-called end-to-end structure, realizes the "encoder-quantizer-decoder" networks, where the encoder compresses the speech signal and extracts the low-dimensional latent features, and the decoder recovers the signal from the quantized features [14-15].

Among neural-based methods, there are two problems. Firstly, the absolute speech quality is not as good as the performance of SP-based methods at high bitrates. Secondly, quantized noise and spectral artifacts are produced when we reduce the complexity of generative models for real-world use.

To resolve the above problems, we propose the Penguins codec with a multi-mode architecture in which the WB core is processed by a novel encoder-quantization-decoder model and the SWB part is processed by a bandwidth extension approach. The Penguins codec incorporates a novel multiscale end-to-end structure combining the light-weight deep generative networks

10.21437/Interspeech.2023-1490

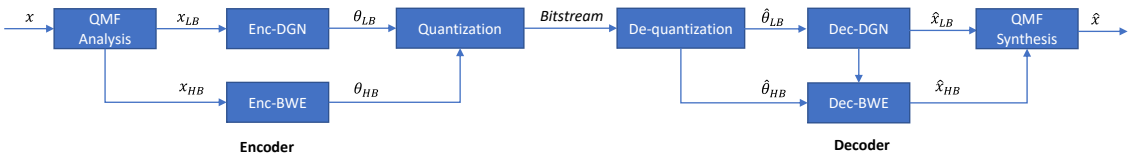

Figure 1: *Flowchart of the Penguins codec.*

with perceptual loss functions and knowledge distillation strategies to efficiently preserve the quality of the reconstructed speech. Additional post-filters are also applied both in the low band and the high band to improve the quality. The subjective and objective experiments prove that the method proposed herein can provide high quality at very low bit rates with moderate computing consumption.

The following paragraphs are organized as follows: We introduce the methods in Section 2. Then we describe the training procedures and strategies in Section 3. We present the subjective and objective experiments including the ablation study in Section 4. The complexity analysis is discussed in Section 5. Finally, the conclusions are depicted in Section 6.

2. Method

In this section, we describe the whole process of the Penguins codec step by step. As illustrated in Fig. 1, the scheme is divided into two parts: Encoder and Decoder. At the Encoder, the input SWB signal is processed by a QMF analysis filter to generate the low band and high band signals; the low band signal is then processed by proposed deep generative networks (Enc-DGN) to extract the feature vector at the low band, while the high band signal is processed by a bandwidth extension encoder (Enc-BWE) to extract the feature vectors at the high band. Both feature vectors in the low band and the high band are quantized and coded into the bitstream. The Decoder is the inverse procedure of the Encoder in which the received feature vectors are sent to the deep generative networks (Dec-DGN) and bandwidth extension decoder (Dec-BWE) to reconstruct the low band and high band signals, respectively. Finally, the QMF synthesis filter is used to reconstruct the final signal. We will introduce the whole procedure in the following paragraphs.

2.1. Encoder of the Penguins codec

The SWB input, denoted by x, is a 16-bit PCM sampled at 32000Hz. Since the frame length is set to 20ms, the number of bins of x is 640.

In this paper, we refer to the so-called 2-band QMF filter [16] to split the SWB input into two sub-band signals, x_{LB} and x_{HB}, and the number of bins of x_{LB} and x_{HB} is 320.

2.1.1. Enc-DGN at low band

As mentioned above, a multi-mode scheme is proposed herein by considering the redundancy of the speech from low frequency to high frequency. Due to the importance of the WB components, we design a block (denoted by Enc-DGN) properly based on a light-weight deep generative network and apply it to the low band signal, x_{LB}, to extract a low-dimensional feature vector at low band, θ_{LB}.

As illustrated in Figure 2, the Enc-DGN is composed of 2 convolution layers, a pre-processing layer, and 4 layers of EncBlock with specific structure.

For each frame of 320 samples, the first convolution layer extracts the internal information and outputs multi-channel features. In this paper, the number of channels of the Enc-DGN is set to 16.

The pre-processing layer includes a casual convolution layer following a ReLU activation and average pooling by a downsampling factor of 2. It is noted that the pre-processing layer does not change the number of channels (i.e., 16).

There are four EncBlock, and each EncBlock is composed of three dilated residual units with a dilation rate $d = \{1,3,9\}$, and an average pooling by a pre-defined downsampling factor. The kernel size of all convolutional layers is 3. Each EncBlock increases the number of channels by 2; therefore, the number of the output channel of the 4th EncBlock is 256. The downsampling factor for the above four EncBlock is $= \{2,4,4,5\}$, resulting in a 256×1 feature expression.

Finally, a convolution layer with $tanh\,(\cdot)$ activate function is applied to obtain an M-dimensional feature vector, θ_{LB}, where M is the final number of channels. The θ_{LB} is in [-1.0, 1,0], and the value of M is dependent on the target bitrate.

2.1.2. Enc-BWE at high band

As the second part of multi-mode coding, we refer to traditional frequency domain BWE (FD-BWE) to generate an artificial high band component. The general idea of FD-BWE is to replicate the selected spectrum patches from the low bands to the high band; the local spectrum envelopes are calculated and transmitted to the decoder to adjust the spectrum in the high band.

In this paper, we transform the x_{HB} to the MDCT (Modified Discrete Cosine Transform) domain [17]. Then, we divide the SWB spectrum into eight sub-bands and calculate the average power in each sub-band in the high frequency ranges as the spectrum envelopes. Finally, we convert the spectrum envelopes into the logarithm domain, denoted by θ_{HB}.

2.1.3. Quantization and entropy coding

Each component in θ_{LB} is scalar quantized by searching the nearest label in pre-defined codebooks with 11 embeddings uniformly distributed between -1.0 and 1.0. The number of the channel of θ_{LB}, denoted by M, is determined by the assigned bitrate. Since the theoretical number of bits in each frame is $(-1 * M * \log 2\left(\frac{1}{11}\right))$, we set the M by 34 to make the bitrate less than 6 kbps for WB core. Then, we implement the entropy coding by preparing the probability distribution function of the embedding in each channel, separately.

Meantime, each of the eight logarithm spectrum envelopes in the high band, θ_{HB}, is quantized with a codebook with 32 components and coded in 5 bits directly. Therefore, the rate for the high band is 2 kbps. Consequently, the total rate of the SWB coding could be less than 8 kbps in the proposed method.

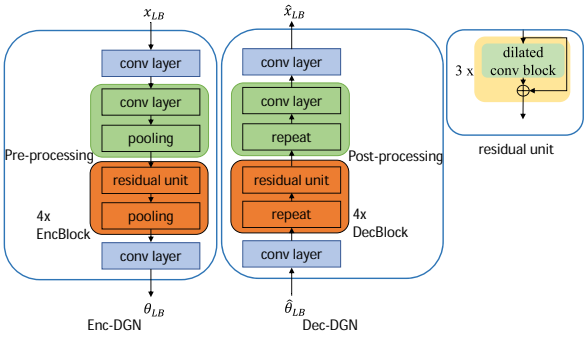

Figure 2: *Structure of Enc-DGN and Dec-DGN.*

2.2. Decoder of the Penguins codec

For each frame, we extract all parameters from the bitstream and obtain the quantized parameters, i.e., $\hat{\theta}_{LB}$ and $\hat{\theta}_{HB}$. Then, we reconstruct the sub-band signals, respectively.

2.2.1. Dec-DGN at low band

As illustrated in Figure 2, the Dec-DGN is composed of 2 convolution layers, 4 layers of DecBlock, and a post-processing layer. The Dec-DGN is a mirror version of the Enc-DGN with similar structures. To reduce the complexity of the decoder, we apply the repeat operations in upsampling layers of Dec-DGN, which is different from the so-called transpose convolution commonly applied in previous works.

Additionally, a post-filter consisting of a 10th-order pole-zero filter and a first-order all-zero filter is applied in the inference phase to strengthen the formant structures of the speech [18]. The transfer function is defined by

$$H_{harmonic}(z) = \frac{1-\sum_{i=0}^{N} b_i z^{-i}}{1-\sum_{i=0}^{N} a_i z^{-i}} * (1+\mu z^{-1}), \qquad (1)$$

where μ is 0.15.

2.2.2. Dec-BWE at high band

Given the reconstructed low band signal by Dec-DGN, $\hat{x}_{LB}$, we implement the MDCT transform to obtain the spectrum in the low band. Then, we calculate the spectrum envelopes in the low band to obtain the normalized MDCT coefficients.

The normalized MDCT coefficients in the low band are replicated to generate artificial coefficients in the high band. Then, eight spectrum envelopes extracted from the bitstream, $\hat{\theta}_{HB}$, are applied to adjust the energy in the high band.

To avoid musical noise from the extra harmonic opponents replicated from the low band, a spectrum whitening process is added. Additionally, a 2-order IIR slope filter is applied to attenuate the spectrum in the SWB range, which is used to improve the subjective experience. The parameters of the IIR filter are listed in Table 1.

Then, the inversed MDCT transform is applied to obtain the high band signal, $\hat{x}_{HB}$.

Finally, a QMF synthesis filter is implemented to generate the speech at 32000 Hz sampling rate, $\hat{x}$, by referring to the reconstructed sub-band signals, $\hat{x}_{LB}$ and $\hat{x}_{HB}$.

Table 1: *Parameters of 2-order IIR slope filter.*

	Paramters
Numerator	0.5690, 1.1381, 0.5690
Denominator	1.0000, 0.9428, 0.3333

3. Dataset and training strategies

3.1. Training dataset

The clean speech in the training dataset is mainly derived from LibriTTS [19], DNS Challenge [20] and private datasets. In addition, we also add some noise speech from DNS Challenge, and we also refer to MIR-1k [21] and FMA[22] datasets to add the music utterances in the training dataset. It should be noted that we only trained the model with English and Mandarin utterances. All utterances are resampled at 16 kHz.

3.2. Training strategies

The generator-discriminator strategy is adopted in our training procedure. The generator is composed of the Enc-DGN and Dec-DGN structures, and the weights of the generator are updated according to the training loss function that will be introduced in Section 3.3.

Our adversarial training framework is based on the multi-resolution STFT-based (MR-STFT) discriminators with 6 different scales with FFT points of {60, 120, 240, 480, 960, 1920}. Each discriminator, composed of seven 2D convolution layers with a kernel size of {3, 3}, takes the logarithmic magnitude spectrum, and all the spectrums are concatenated as the input. Weight normalization and LeakyReLU are applied sequentially after each 2D convolution layer except the last one.

As mentioned before, some low-complexity factors are considered both in Enc-DGN and Dec-BGN. To further improve speech quality, we also append an additional GAN-knowledge distillation (KD) algorithm after the adversarial training stage, denoted by GAN-KD. We pre-train a high-computational complexity model as the teacher, $\{G_T; D_T\}$, which transfers the information to the proposed model (student model), $\{G_S; D_S\}$ with GAN-KD. During the distillation stage, we calculate the distillation loss by evaluating the distance between the intermediate outputs of the teacher model and the student model.

3.3. Training loss functions

The training loss functions at the GAN stage include prediction loss, adversarial loss, and feature match loss.

The prediction loss is used to evaluate the loss of the reconstructed spectrum and is composed of a multi-resolution short-time Fourier transform (STFT) loss and a perceptual constraint loss. For the multi-resolution STFT loss, we try to minimize the spectral convergence loss and the L1 distance in the logarithmic magnitude spectral domain as below

$$\mathcal{L}_s(X) = \sum_r \left\| \log(X_r) - \log(\hat{X}_r) \right\|_1 + \frac{\|X_r - \hat{X}_r\|_F}{\|\hat{X}_r\|_F}, \qquad (2)$$

As the second part of prediction loss, some perceptual constraints are adopted to preserve the accuracy of the valleys of the harmonics in the predicted speech (like the perceptual weighting in the CELP scheme). In this paper, we convert the STFT spectrum of the predicted and target signal into the equivalent rectangular bandwidth (ERB) [23] and calculate the average energy of each ERB. Then the perceptual loss is obtained by

$$\mathcal{L}_{PE}(X) = \left\| P_{ERB}(X) - P_{ERB}(\hat{X}) \right\|_1, \qquad (3)$$

where the $P_{ERB}(X)$ and $P_{ERB}(\hat{X})$ are the average energies in the ERB of the target and predicted signal.

In addition, we also append the adversarial loss, $\mathcal{L}_{adv}$, and

feature match loss, $\mathcal{L}_{fe}$, which are commonly used in previous works.

The overall generator loss is a weighted sum of the above loss terms

$$\mathcal{L}_G = \lambda_S \cdot \mathcal{L}_S + \lambda_{PE} \cdot \mathcal{L}_{PE} + \lambda_{adv} \cdot \mathcal{L}_{adv} + \lambda_{fe} \cdot \mathcal{L}_{fe}, \quad (4)$$

The training loss function at GAN-KD stage is defined by

$$\mathcal{L}_{KD} = \lambda_G \sum_i \left\| O_i^{G_T} - O_i^{G_T} \right\|_1 + \lambda_D \sum_i \left\| O_i^{D_T} - O_i^{D_T} \right\|_1, \quad (5)$$

where O_i^{Θ} is the i-th layer of the model $\Theta \in \{G_T; D_T; G_S; D_S\}$.

3.4. Training configurations

We train the models for 2000000 steps with the AdamW optimizer and the ExponentialLR scheduler. The batch size is set to 16. Each clip is randomly selected for 2 seconds for training.

Empirically, the weights $\{\lambda_S, \lambda_{PE}, \lambda_{adv}, \lambda_{fe}, \lambda_G, \lambda_D\}$ are set to {1, 2, 1, 20, 30, 10}.

4. Evaluations and discussions

4.1. Objective evaluation

We refer to the ITU P.863 [24] as the objective evaluation metric. Given the reference and degraded signals, the P.863 outputs a predicted Mean Opinion Score (MOS). As the successor of PESQ [25], the P.863 provides two operational modes, in which the SWB mode supports quality evaluation of WB and SWB speech on a single scale, and the MOS score predicted is between 1.0 and 4.75.

The test set is from the ITU-T P.501 database [26] which is not used for training the proposed method. There are eight different languages, and each language contains two male and two female utterances; therefore, 32 utterances are selected.

We employ the OPUS codec and state-of-the-art open-source neural codecs (i.e., Lyra2 [27] and Encodec [28]) as the anchor systems to evaluate the performance of the proposed method by comparing the average MOS by P.863 SWB mode.

We illustrate the P.863 SWB scores in Figure 3, including scores for WB codecs (blue) and SWB codecs (green) at different bitrates. We observe that the proposed Penguins codec outperforms OPUS and neural codecs within similar bitrates, not only in WB but also SWB parts. Among the three systems related to the proposed methods, adding the perceptual constraint loss function (*6kbps_wiPE) and GAN-KD (*6kbps_wiPE&KD) can obtain incremental improvement.

4.2. Subjective listening test

We also perform a crowdsourced subjective listening test according to the ITU-T P.808 recommendation [29]. We divide the listening test into WB and SWB parts, respectively. Both parts refer to the same SWB clean reference speech utterances, and the Degradation Category Rating (DCR) is selected to rank the quality. There are 12 speech utterances in Chinese, and we invite 24 native listeners to participate in the listening test.

The listening test results are illustrated in Figure 4. The subjective quality of proposed Penguins codec is better than OPUS and other neural codecs at low bitrates, which is consistent with objective evaluation results. Furthermore, the subjective MOS of the Penguins codec is also comparable to that of OPUS at high bitrates, which demonstrates the excellent performance of the Penguins codec.

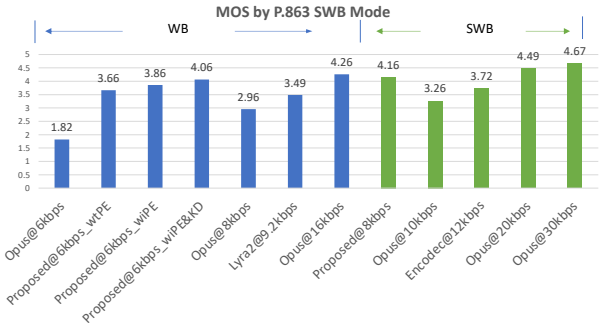

Figure 3: *Objective evaluation results.*

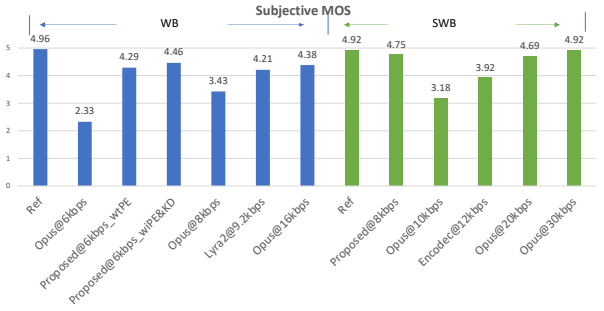

Figure 4: *Subjective listening test results.*

5. Complexity analysis

We compare the real-time factor (RTF) over different neural codecs according to a single thread implementation in the MacBooPro 2019 (i7 2.6GHz), and the RTF results are listed in Table 2. It is observed that the Penguins codec proposed in this paper is better than Encodec in terms of RTF. Meanwhile, the complexity of Penguins codec is also comparable to Lyra2. Therefore, we can conclude that the proposed method in this paper is sufficient to fulfill the real-time services.

Table 2: *Real time factor (RTF) of neural codecs.*

	Encoder	Decoder
Lyra2@9.2kbps	0.015	0.034
Encodec@12kbps	0.103	0.094
Penguins@6kbps	0.032	0.028

6. Conclusions

In this paper, we propose the Penguins codec with a multi-mode neural speech coding scheme with high-quality reconstruction. It includes a WB core based on encoder-decoder structures with deep generative networks and SWB reconstruction based on frequency domain bandwidth extension. Especially, we design the training strategies properly with proposed perceptual constraint loss and knowledge distillations to improve the performance of the GAN architecture. Both objective and subjective experiments prove the method proposed in this paper can outperform traditional and neural codecs at very low bitrates and be comparable to the traditional codec at high bitrates. Especially, we provide the ablation experiments to verify the benefits of different strategies adopted by the Penguins codec. The empirical results in computing efficiency also indicate the Penguins codec can fulfill the requirement for real-time communications. Future work includes further quality improvement for diverse RTC applications by investigating novel deep neural networks and perceptual loss controls.

7. References

[1] ITU-T Recommendation G.711: *Pulse code modulation (PCM) of voice frequencies*. 1972.

[2] ITU-T Recommendation G.722: *7kHz audio-coding within 64kbps*. 1988.

[3] M. Schroeder and B. Atal, "Code-excited linear prediction (CELP): high-quality speech at very low bit rates," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1985, pp. 937–940.

[4] 3GPP TS26.171: *Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; General description*. 2002.

[5] 3GPP TS26.441: *Codec for Enhanced Voice Services (EVS); General overview*. 2014.

[6] IETF RFC6716: *Definition of the opus audio codec*. 2012.

[7] T. Painter and A. Spanias, "Perceptual coding of digital audio", in *Proceeding of the IEEE*, 2000, pp. 452-513.

[8] ITU-T Recommendation G.729.1: *G.729-based embedded variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729*. 2006.

[9] B. Geiser, P. Jax, P. Vary, H. Taddei, S. Schandl, M. Gartner, C. Guillaume and S. Ragot, "Bandwidth extension for hierarchical speech and audio coding in ITU-T Rec. G.729.1," in *IEEE Transaction on audio, speech, and language processing*. IEEE, 2007, pp. 2496-2509.

[10] P. Ekstrand: "Bandwidth Extension of Audio Signals by Spectral Band Replication", *Proc.1st IEEE Benelux Workshop on Model based Processing and Coding of Audio (MPCA-2002)*, Leuven, Belgium, November 15, 2002.

[11] W. B. Kleijn, F. Lim, A. Luebs, J. Skoglund, F. Stimberg, Q. Wang and T. Walters, "WaveNet based low rate speech coding," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018, pp. 676–680.

[12] J.-M. Valin and J. Skoglund, "Lpcnet: Improving neural speech synthesis through linear prediction," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5891–5895.

[13] https://github.com/google/lyra/releases

[14] sN. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[15] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[16] P. P. Vaidyanathan, "*Multirate systems and filter banks*", Pearson College Div, 1992.

[17] H. Malvar, "Lapped Transforms for Efficient Transform / Subband Coding", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 1990, pp. 969–978.

[18] ITU-T Recommendation G.728: *Coding of speech at 16kbit/s using low-delay code excited linear prediction*. 1992.

[19] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[20] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun et al., "The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," *arXiv preprint arXiv:2005.13981*, 2020.

[21] C.-L. Hsu and J.-S. R. Jang, "On the improvement of singing voice separation for monaural recordings using the mir-1k dataset," *IEEE transactions on audio, speech, and language processing*, vol. 18, no. 2, pp. 310–319, 2009.

[22] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "Fma: A dataset for music analysis," *arXiv preprint arXiv:1612.01840*, 2016.

[23] B. Glasberg and B. Moore. "Derivation of Auditory Filter Shapes from Notched-Noise Data." *Hearing Research*. Vol. 47, Issues 1–2, 1990, pp. 103–138.

[24] ITU-T Recommendation P.863: *Perceptual objective listening quality prediction*. 2011.

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

[26] ITU-T Recommendation P.501: *Test signals for use in telephony and other speech-based applications*. 2020.

[27] https://opensource.googleblog.com/2022/09/lyra-v2-a-better-faster-and-more-versatile-speech-codec.html

[28] https://github.com/facebookresearch/encodec

[29] ITU-T Recommendation P.808: *Subjective evaluation of speech quality with a crowdsourcing approach*. 2021.