



# LSPnet: an ultra-low bitrate hybrid neural codec

Bowen Zhang<sup>1,2</sup>, Ian McLoughlin<sup>1</sup>, Xiaoxiao Miao<sup>1,3</sup>, AS Madhukumar<sup>2</sup>

<sup>1</sup>ICT Cluster, Singapore Institute of Technology, Singapore

<sup>2</sup>College of Computing & Data Science, Nanyang Technological University, Singapore

<sup>3</sup>Division of natural and applied sciences, Duke Kunshan University, China

bowen009@e.ntu.edu.sg, ian.mcloughlin@singaporetech.edu.sg,  
xiaoxiao.miao@singaporetech.edu.sg, ASMadhukumar@ntu.edu.sg

## Abstract

This paper presents an ultra-low bitrate speech codec that achieves high-fidelity speech coding at 1.2kbps while maintaining low computational complexity. Building upon the LPCNet framework, combined with a parametric encoder, we introduce several key improvements by incorporating line spectral pairs (LSP) to improve quantization error performance and eliminate explicit LPC estimation by directly predicting the probability distribution of audio samples using a deep neural network, and employing a joint time-frequency training strategy combining short-time Fourier transform (STFT) loss with cross-entropy (CE) loss. The codec is suitable for real-time applications in resource-constrained environments. Experimental results show that the proposed codec not only outperforms traditional speech codecs but also achieves superior speech quality compared to state-of-the-art end-to-end codecs, offering a compelling balance between quality and computational cost.

**Index Terms:** hybrid speech codec, neural vocoder

## 1. Introduction

Ultra-low bitrate speech coding is important for highly bandwidth-limited communications systems, such as military radio, satellite links, and even voice transmission over IoT networks. However, achieving intelligible and natural-sounding speech at bitrates as low as 1.2 kbps is difficult. Traditional speech codecs, such as MELP (Mixed-Excitation Linear Prediction) [1] and CELP (Code-Excited Linear Prediction) [2], rely on compact signal representations and highly efficient quantization strategies, to preserve naturalness and speaker identity, but this deteriorates significantly at very low bitrates. The result is often robotic speech or muffled artifacts due to the aggressive compression applied to speech parameters [3].

To address such limitations, deep neural network (DNN)-based speech coding approaches have emerged. These demonstrate significant improvements in quality. Initial hybrid methods, such as RNNoise [4] for low-bitrate noise suppression and WaveRNN-based codecs [5], introduced the potential of combining deep neural networks with traditional parametric approaches. For hybrid neural codecs, the decoder input can be a direct bitstream [6] or the quantized speech features [7, 8]. The most famous hybrid method, LPCNet [7], combines a deep neural vocoder and a parametric codec. It leverages neural networks to improve the speech synthesis aspects, while maintaining a relatively low computational footprint. However, it inherits certain limitations from the underlying parametric frameworks, including the sensitivity of LPC (Linear Predictive Coding) coefficients to quantization errors.

Beyond hybrid methods, fully end-to-end deep learning-based speech coding frameworks, such as SoundStream [9], En-

Codec [10], and DAC [11], have demonstrated that neural networks can effectively encode and reconstruct speech without using explicit parametric feature extraction. These approaches rely on learned representations and powerful generative models, such as convolutional neural networks or transformers, to encode speech into a compact latent space and reconstruct it with good fidelity. These methods can achieve remarkable quality at moderate to low bitrates, but the complexity is much higher than conventional codecs, thus impractical for resource-constrained applications. Decoder complexity in some end-to-end approaches uses autoregressive or diffusion-based synthesis, with significant latency as well as complexity.

We propose LSPnet, an ultra-low bitrate speech codec aiming for high quality speech at 1.2 kbps with low computational cost. Our method builds upon LPCNet, introducing several key improvements to enhance both quality and efficiency: (a) We convey spectral envelope via line spectral pairs (LSPs), which have greater numerical stability and tolerance to quantization errors than BFCC (Bark frequency cepstral coefficients), leading to improved spectral accuracy. (b) We replace LPCNet's explicit linear prediction estimation stage with direct audio sample prediction. This eliminates intermediate filter coefficient representations like BFCC and LPC, reduces conversions, and avoids the need to force the network to implicitly compensate multiple error sources in excitation modeling. (c) To further enhance speech quality, we incorporate a joint training strategy using short-time Fourier transform (STFT)-based loss and cross-entropy (CE) loss to balance frame-wise perceptual error against instantaneous statistical accuracy. (d) To reduce complexity, we use an efficient parametric encoder to extract speech features. This not only lowers the encoding complexity but enables us to build a lightweight decoder, which is important for low-resource deployments. A hybrid of neural-based synthesis, low resource feature extraction, and efficient encoding, enables an effective solution for ultra-low bitrate speech coding, balancing high-quality speech output with real-time operability. We compare against existing ultra-low bitrate codecs including both state-of-the-art systems as well as vanilla LPCNet, demonstrating significant improvements in quality.

## 2. Related Work

A typical speech codec consists of encoder-decoder and quantizer-dequantizer pairs. At the transmitter end, an encoder extracts parameters or features from input speech, while at the receiver end, a decoder reconstructs the speech signal from its internal history and the received features. A quantizer at the transmitter end, matched to a dequantizer at the receiver end, aim to reduce the transmission size of the parameters or features, while minimising quality degradation due to

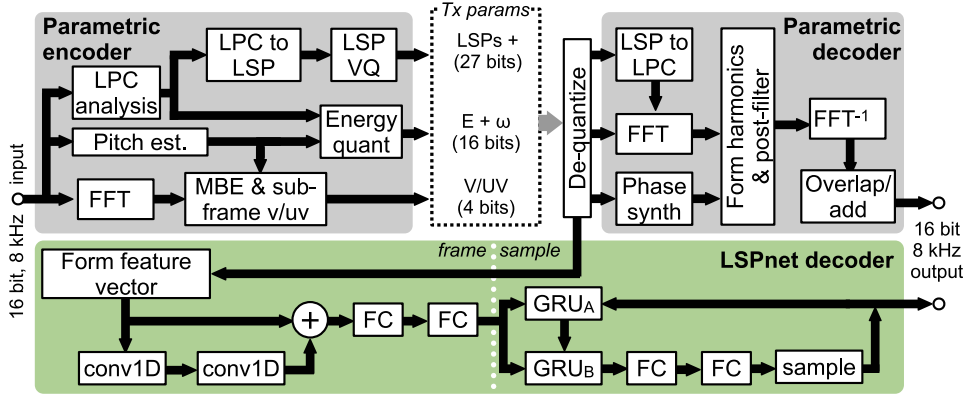


Figure 1: Architecture of the LSPnet hybrid parametric-neural codec, showing the Codec-2 encoder (top left) and original decoder (top right). The LSPnet decoder operates on the transmission bitstream at frame level (bottom left) and sample level (bottom right).

quantization error. Early codecs such as G.711 [12], used simple pulse code modulation and logarithmic scaling to achieve voice-quality coding at 8kHz, with a relatively high bitrate of 64kbps. Such codecs have generally been replaced with algorithms that achieve higher quality at lower bitrate, thanks to advanced computational techniques. G.729 and AMR [13] both employ CELP-based structures, while more recent codecs such as Opus [14] and EVS [15] incorporate more sophisticated signal-processing techniques.

Low bitrate codecs often make use of linear predictive coding (LPC) and pitch filtering techniques. LPCs eliminate redundancy by predicting a frame of future samples based on a linear speech model applied to past samples. Pitch filtering handles periodicities in the signal that can extend beyond frames. Such systems generally prioritise intelligibility and often struggle to maintain quality at very low bitrates [16]. Opus, notable, is generally considered to maintain good speech quality even at low bitrates, while also offering flexible bitrate control.

End-to-end neural codecs jointly optimize both encoding and decoding using deep neural networks. This approach enables more flexibility in the compressed representation, which learns directly from waveform data. Models such as [9, 10, 11, 17] employ end-to-end GAN-based encoder-decoder structures, utilizing quantization methods such as residual vector quantization (RVQ) [9], and variants such as group RVQ (GRVQ) [11]. These models generally achieve higher perceptual quality at low bitrates, but at the cost of considerable computational power. Their complexity limits deployment in real-time or resource-constrained environments.

Line spectral pairs (LSP) [18], also known as line spectral frequencies (LSF), are vocal tract resonance representations that convey largely the same information as the more common BFCC of the same order, but with significant improvement in quantizability and stability [19, 20]. For these reasons, they have been used in several advanced codecs, including Opus [14] and Codec2 [21, 22], an open source parametric speech compressor. LSPs are also amenable to perceptual modification [23].

### 3. Proposed Method

In this paper we propose LSPnet, a neural decoder adopting a hybrid approach inspired by LPCNet [7]. LSPnet leverages the low complexity encoder part of Codec2 [22], to extract essential features to represent encoded audio, quantize them, and en-

code them into a compact bitstream. We then replace the function of the Codec2 decoder by training a neural decoder with the dequantized bitstream as feature vector, and an objective to enhance reconstructed speech quality. The aim is to maintain the low complexity of Codec2 at the encoder end while taking advantage of the quality enhancement properties of deep neural synthesis with relatively low complexity at the decoder end. The structure of the proposed LSPnet is shown in Fig. 1, and discussed in the following subsections. At present, LSPnet’s transmission bitrate is fixed to 1.2kbps, targeting ultra-low bitrate applications.

#### 3.1. Parametric Encoder

The purpose of the encoder is to extract the essential features required for reconstructing speech while effectively compressing the information. A neural architecture could be employed here to train an encoder that learns transmission parameters in a data-driven manner, such as in end-to-end neural approaches [9, 11]. However, such approaches typically demand high levels of computational resource, potentially increasing latency and complexity, making them less practical for real-time or resource-constrained applications. Instead, as aforementioned, we leverage an efficient parametric encoder that explicitly models the human vocal system, yielding features such as spectral envelope (reflecting vocal tract resonance), pitch (vocal fold vibration), and voicing level (indicating phonation characteristics) [16].

The Codec2 sinusoidal coder in 1.2kbps mode is configured to operate on 40ms frames at a sample rate of 8kHz. The encoder extracts 10th order LSPs, energy, pitch, and voiced/unvoiced flags to represent speech. These parameters are compressed to a bitstream for transmission with each 40ms frame being encoded to 48 bits as shown in Fig. 1, with 1 bit per frame reserved. Multi-stage vector quantization (VQ) is used for the LSPs [24] while joint VQ is used for signal energy and pitch. Four bits indicate voicing level [25] per subframe.

#### 3.2. Neural Decoder

The input features we use to train LSPnet are extracted from the dequantized Codec2 bitstream. Unlike LPCNet, which uses 18 cepstral coefficients computed on the Bark scale as input, we rely largely on 10th order LSPs for spectral envelope information.

The architecture of LSPnet, shown in Fig. 1, inherits the dual rate structure of LPCNet with frame-by-frame and sam-

ple rate modules. At the frame level, input frames are processed through a network comprising two convolutional layers followed by two dense layers. These extract latent frame-level features to capture temporal dependencies and high-level patterns from the audio signal. At the sample level, the frame-level latent features are further passed through two GRU layers and a dual dense layer which predict the probability distribution of individual audio samples. Unlike LPCNet, which predicts the residual difference between the actual sample value and the linear prediction, LSPnet directly predicts sample value in the  $\mu$ -law domain before conversion to 16-bit linear PCM. This direct prediction approach simplifies the training objective and eliminates reliance on intermediate LPC computations (i.e. we avoid the BFCC-LPC conversion as well as autoregressive linear prediction steps), allowing a streamlined process.

### 3.3. Loss Function

Following initial empirical evaluations, LSPnet was trained to minimise a weighted loss that balances a frame-level STFT loss component similar to FARGAN but excluding its feature matching loss [26], and sample-wise cross-entropy loss, as used effectively in the original LPCNet [7].

**STFT Loss:** This term compares the model output  $s' = LSPnet\{Codec2\_enc(s)\}$  from input frame  $s$ , to the ground truth input in the frequency domain. Hence it compares generated spectral response  $S' = FFT(s')$  against Hamming windowed  $w$  input spectral response  $S = FFT(s \times w)$ ,

$$\mathcal{L}_S = \sum_{t,f} \left[ (|S_{t,f}| - |S'_{t,f}|)^2 + 0.1 |\angle S_{t,f} - \angle S'_{t,f}| \right] \quad (1)$$

**CE Loss:** This operates in the time domain to optimize the sparse categorical cross-entropy loss between the predicted  $s'$  and ground truth  $s$  sample distributions over  $N$  samples and  $J$  classes,

$$\mathcal{L}_{CE} = - \sum_{n=1}^N \log \left( e^{s'_n} / \sum_{j=1}^J e^{s'_j} \right) \quad (2)$$

Overall loss  $\mathcal{L}_{tot}$  balances time and spectral components via factor  $\alpha$  to ensure both sample and frame-level accuracy in the output speech. We investigate the influence of  $\alpha$  in section 4.2.1,

$$\mathcal{L}_{tot} = \alpha \mathcal{L}_S + (1 - \alpha) \mathcal{L}_{CE} \quad (3)$$

### 3.4. Training Strategy

To improve model robustness and performance, a pseudo-batch training approach was adopted. During training, the frames within each batch are shuffled randomly to ensure a diversity of frame types, which we found produced superior results compared to correlated batches. Additionally, to increase the model's resilience to noise and improve generalization, discrete uniform distributed random noise is introduced in the  $\mu$ -law domain of the audio during training. This augmentation technique helps the model adapt to real-world conditions, ultimately enhancing its reliability and performance in diverse scenarios.

## 4. Evaluation

In this section, we first optimize the loss function by training the model with different STFT and CE loss weight balancer  $\alpha$  on a subset of LibriTTS [27]. We further scale up training on the full LibriTTS dataset with pseudo-batch processing. Both objective and subjective evaluation of our proposed model was done, alongside comparison to different existing codecs.

### 4.1. Experiment Setup

Training and testing used separate datasets: LibriTTS [27] for training and TIMIT [28] for testing, to make the results more convincing for real-world application. During training, the speech utterances are first peak normalized and resampled to 8kHz, then concatenated to form the raw speech. Concatenated raw speech frames of 80 samples (10ms) are chunked into 15 continuous frames, and 48 chunks at a time are randomly batched (shuffled) like in LPCNet. During testing, 56 recordings are randomly selected from TIMIT-test with 6 of the sentences (3 males and 3 females) reserved for subjective listening tests, and 50 sentences reserved for objective evaluation.

### 4.2. Objective Evaluation

While listening tests are ideal for comparing speech quality, objective evaluation is much easier to conduct on a large scale. STOI (Short-Time Objective Intelligibility) and PESQ (Perceptual Evaluation of Speech Quality) are two widely used objective metrics for evaluating speech intelligibility and quality. We utilized open source toolkits<sup>12</sup> to calculate both. For codecs operating at different sample rates, PESQ was set to 'nb' for 8kHz speech and 'wb' for 16kHz and 24kHz audio.

#### 4.2.1. Determining optimal loss weighting

To determine the optimal loss weight  $\alpha$  efficiently, we trained the model with a subset of around 3.5h (i.e. 2000 utterances) randomly sampled from *LibriTTS-train-clean-100* for  $\alpha$  between 0.1 and 0.9. The results in Table 1 interestingly reveal that an even balance, giving equal importance to both time and frequency domain loss, produces the highest average objective scores. Hence we fix  $\alpha = 0.5$  for the remaining tests.

Table 1: Effect of different  $\alpha$  on objective quality.

$\alpha$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
<b>STOI</b>	0.91	0.91	0.92	<b>0.93</b>	0.92	0.92	0.92	0.91	0.91
<b>PESQ</b>	2.17	2.68	2.74	2.80	<b>2.84</b>	2.82	2.58	2.41	2.35

#### 4.2.2. Objective Quality Comparison

Based on optimal loss weighting and further full dataset training, we compare LSPnet to several publicly available traditional and modern neural codecs. All were tested on the same 50 randomly selected samples from TIMIT-test. For the codecs operating at different sample rates to the 16kHz source audio, we resampled the testset and normalized gain using sox. Opus<sup>3</sup>, Encodec<sup>4</sup> and DAC<sup>5</sup> are optimized for wideband samples at higher bitrates, but they also support a lower bitrate mode, with degraded audio quality. Codec2 results are for the original parametric decoder and encoder.

A summary of results in Table 2 demonstrates the performance of the evaluated codecs for different bitrates and sample rates. The proposed LSPnet, operating at 1.2 kbps achieves a higher average STOI and PESQ to similar bitrate codecs, outperforming both Codec2 and LPCNet. Higher bitrate codecs like Opus and Encodec can improve PESQ but at a significantly higher bitrate cost. DAC@5.33 kbps achieves the best PESQ

<sup>1</sup><https://pypi.org/project/pesq/>

<sup>2</sup><https://github.com/mmpariente/pystoi>

<sup>3</sup><https://github.com/xiph/opus>

<sup>4</sup><https://github.com/facebookresearch/encodec?tab=readme-ov-file>

<sup>5</sup><https://github.com/descriptinc/descript-audio-codec/tree/main>

Table 2: Model complexity and objective quality comparison.

Codec	Bitrate	Sample rate	Params	STOI	PESQ
Opus [14]	3kbps	16kHz	-	0.67	1.48
	6kbps	16kHz	-	0.86	2.56
Codec2 [29]	1.2kbps	8kHz	-	0.83	2.64
	2.4kbps	8kHz	-	0.86	2.98
LPCNet [7]	1.6kbps	16kHz	1.26M	0.65	1.40
Encodec [10]	1.5kbps	24kHz	-	0.79	1.61
	3kbps	24kHz	14.85M	0.84	2.06
	6kbps	24kHz	-	0.87	2.60
	1.78kbps	16kHz	-	0.79	1.80
DAC [30]	2.76kbps	16kHz	74.18M	0.83	2.42
	5.33kbps	16kHz	-	0.87	3.48
Proposed	1.2kbps	8kHz	1.24M	0.94	3.27

Table 3: Two-alternative forced choice results.

codec	preference (%)	codec
reference	86.7 ← 13.3	LPCNet 1.6kbps
reference	93.3 ← 6.7	Codec2 1.2kbps
reference	66.7 ← 33.3	LSPnet 1.2kbps
Codec2 1.2kbps	10.8 → 89.2	LSPnet 1.6kbps
LPCNet 1.6kbps	15.0 → 85.0	LSPnet 1.2kbps

but requires around 74M parameters, making it computationally expensive. In contrast, our approach maintains a strong balance between efficiency and intelligibility, achieving high speech quality at a fraction of the bitrate and complexity.

Figure 2 compares STOI and PESQ results for each utterance. Codecs having swithcable bitrates perform much better as bitrate increases, but among tested methods, only DAC is capable of outperforming LSPnet in terms of PESQ, at the highest quality 5.33kbps mode – more than 4 times the bitrate, and with over 50 times the number of learnable parameters.

### 4.3. Subjective Evaluation

In order to validate the objective findings, subjective two-alternative forced-choice listening tests were conducted to assess codec quality. The tests involved presenting randomised pairwise comparisons between reference (uncoded) speech, and the speech from Codec2 at 1.2kbps, LPCnet at 1.6kbps and the proposed LSPnet at 1.2kbps, i.e. the nearest like-for-like conditions. In total, 20 individuals aged between 18 and 35, with self-reported normal hearing, were tested.

The results from these listening tests, presented in Table 3, demonstrates a clear preference for the proposed LSPnet over other low-bitrate codecs. For example, LSPnet outperformed Codec2 and LPCNet by a factor of around 8×, indicating easily perceptible superior speech quality. Compared to the uncoded 8kHz reference, LSPnet was chosen 33.3% of the time, showing noticeable, but limited, degradation. Codec2 and LPCNet, by contrast, were rarely preferred over reference speech, highlighting their lower perceptual quality. These findings confirm that LSPnet achieves a significant quality improvement, making it a strong candidate for low-bitrate speech coding applications.

### 4.4. Discussion of results

Firstly we note that LSPnet is very well trained at fixed 1.2kbps bitrate in 8kHz, whereas some of the other codecs have multiple operating modes and may have sacrificed performance for flexibility. Secondly, LSPnet outperforming Codec2 indicates that the neural network architecture is performing enhancement,

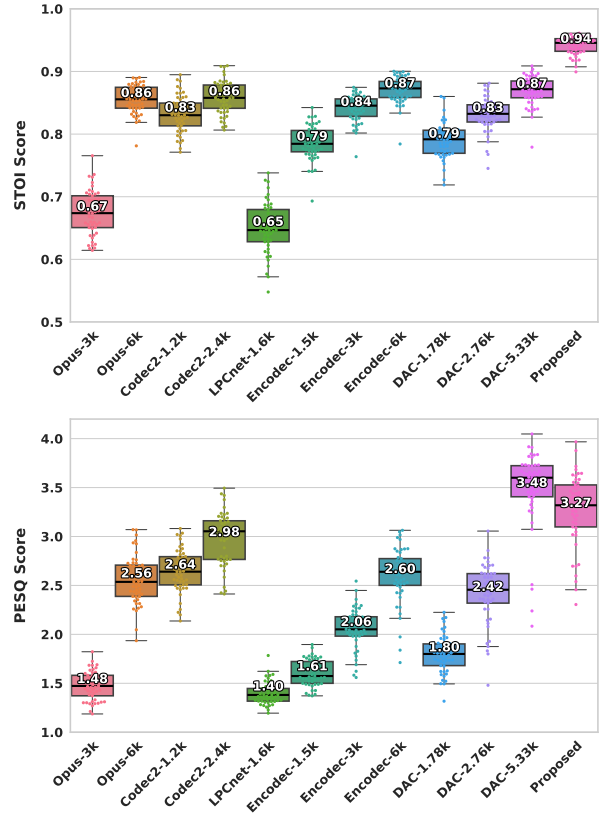


Figure 2: Comparison of utterance-level scores for different codecs in terms of (a) STOI and (b) PESQ. The box plots indicate the mean, upper and lower quartile and spread of results.

as anticipated. LSPs are known to outperform LPCs in highly quantized scenarios [19], such as the ultra-low bitrate mode we are targeting. Hence the significant performance gain of LSPnet over LPCNet is expected. In high bitrate scenarios, some of the advantage gained by using LSPs may disappear. Finally, parametric codecs explicitly model the human vocal system. This mitigates against the effects of quantization error on quality because even erroneous sounds resemble human speech, in contrast to wideband codecs where errors can lead to annoying sounds (e.g. “bloops”, squeaks, clicks).

## 5. Conclusion

This paper has proposed LSPnet, a novel low-complexity neural-hybrid codec architecture for ultra low bitrate operation. LSPnet leverages the low complexity LSP-based Codec2 parametric encoder operating at 1.2kbps, and utilises a deep neural architecture inspired by LPCNet to reconstruct output audio. LSPnet introduces three key improvements to LPCNet which are (a) using line spectral pairs (LSP) to represent spectral envelope, (b) combining sample-wise time-domain loss and frame-wise frequency-domain loss to benefit the training of the frame- and sample-level dual rate regions of LSPnet, (c) Bypassing LPC reconstruction to avoid the need for BFCC-LPC conversion, avoid computing a loss in the LPC or PARCOR domains, and reconstructing samples directly to fit the spectral envelope without LPC synthesis. Both subjective and objective evaluations confirm the superiority of LSPnet over similar codecs and conditions. In future we aim to compare to more codecs, extend to different bitrates and introduce a wideband 16kHz mode.

## 6. References

- [1] A. V. McCree and T. P. Barnwell, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 242–250, 1995.
- [2] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): High-quality speech at very low bit rates," in *ICASSP '85. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 10. IEEE, 1985, pp. 937–940.
- [3] A. S. Spanias, "Speech coding: A tutorial review," *Proceedings of the IEEE*, vol. 82, no. 10, pp. 1541–1582, 2002.
- [4] J.-M. Valin, "A hybrid dsp/deep learning approach to real-time full-band speech enhancement," in *2018 IEEE 20th international workshop on multimedia signal processing (MMSP)*. IEEE, 2018, pp. 1–5.
- [5] N. Kalchbrenner, A. van den Oord, and K. Simonyan, "Efficient neural audio synthesis," in *Proc. International Conference on Learning Representations (ICLR)*, 2018.
- [6] H. Y. Kim, J. W. Yoon, W. I. Cho, and N. S. Kim, "Neurally optimized decoder for low bitrate speech codec," *IEEE Signal Processing Letters*, vol. 29, pp. 244–248, 2021.
- [7] 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.
- [8] W. B. Kleijn, F. S. Lim, A. Luebs, J. Skoglund, F. Stimberg, Q. Wang, and T. C. Walters, "Wavenet based low rate speech coding," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 676–680.
- [9] N. 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.
- [10] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.
- [11] D. Yang, S. Liu, R. Huang, J. Tian, C. Weng, and Y. Zou, "Hifi-codec: Group-residual vector quantization for high fidelity audio codec," *arXiv preprint arXiv:2305.02765*, 2023.
- [12] Y. Hiwasaki, S. Sasaki, H. Ohmuro, T. Mori, J. Seong, M. S. Lee, B. Kövesi, S. Ragot, J.-L. Garcia, C. Marro, L. Miao, J. Xu, V. Malenovsky, J. Lapierre, and R. Lefebvre, "G.711.1: A wide-band extension to ITU-T G.711," in *2008 16th European Signal Processing Conference*, 2008, pp. 1–5.
- [13] E. Ekudden, R. Hagen, I. Johansson, and J. Svedberg, "The adaptive multi-rate speech coder," in *1999 IEEE Workshop on Speech Coding Proceedings. Model, Coders, and Error Criteria (Cat. No. 99EX351)*. IEEE, 1999, pp. 117–119.
- [14] J.-M. Valin, K. Vos, and T. Terriberry, "Definition of the opus audio codec," Tech. Rep., 2012.
- [15] M. Dietz, M. Multrus, V. Eksler, V. Malenovsky, E. Norvell, H. Pobloth, L. Miao, Z. Wang, L. Laaksonen, A. Vasilache *et al.*, "Overview of the evs codec architecture," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5698–5702.
- [16] I. V. McLoughlin, *Speech and Audio Processing: a MATLAB-based approach*. Cambridge University Press, 2016.
- [17] L. Xu, J. Wang, J. Zhang, and X. Xie, "Lightcodec: A high fidelity neural audio codec with low computation complexity," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 586–590.
- [18] I. V. McLoughlin, "Review: Line spectral pairs," *Signal processing*, vol. 88, no. 3, pp. 448–467, 2008.
- [19] I. V. McLoughlin and F. Hui, "Novel dynamic bit allocation method for LSP quantization," *Proc. IEEE Region 10 conference*, no. 79, Sep. 2000.
- [20] M. Bouzid, N. Meziane, and S.-E. Cheraitia, "Multi-coder vector quantizer for transparent coding of wideband speech isf parameters," *International Journal of Speech Technology*, vol. 27, no. 1, pp. 121–132, 2024.
- [21] F. Ben Ali and S. Djaziri-Larbi, "A long term harmonic plus noise model for narrow-band speech coding at very low bit-rates," in *2017 40th International Conference on Telecommunications and Signal Processing (TSP)*, 2017, pp. 372–376.
- [22] D. Rowe, "Techniques for harmonic sinusoidal coding," in *University of South Australia*, 1997. [Online]. Available: <https://api.semanticscholar.org/CorpusID:9016839>
- [23] I. V. McLoughlin and R. Chance, "LSP-based speech modification for intelligibility enhancement," in *Proceedings of 13th International Conference on Digital Signal Processing*, vol. 2. IEEE, 1997, pp. 591–594.
- [24] J.-M. Valin, "The speex codec manual version 1.2 beta 3," *Xiph.org Foundation*, 2007.
- [25] D. W. Griffin and J. S. Lim, "Multiband excitation vocoder," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 36, no. 8, pp. 1223–1235, 1988.
- [26] J.-M. Valin, A. Mustafa *et al.*, "Very low complexity speech synthesis using framewise autoregressive GAN (FARGAN) with pitch prediction," *IEEE Signal Processing Letters*, 2024.
- [27] 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," in *INTERSPEECH 2019*, ser. Interspeech, 2019, pp. 1526–1530.
- [28] J. S. Garofolo, L. F. Lamel, W. M. Fisher, D. S. Pallett, N. L. Dahlgren, V. Zue, and J. G. Fiscus, "Timit acoustic-phonetic continuous speech corpus," (*No Title*), 1993.
- [29] D. Rowe, "Codec 2," 2018, available online: [http://www.rowetel.com/?page\\_id=452](http://www.rowetel.com/?page_id=452).
- [30] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved rvqgan," *Advances in Neural Information Processing Systems*, vol. 36, pp. 27 980–27 993, 2023.