INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

An Efficient Subband Linear Prediction for LPCNet-based Neural Synthesis

Yang Cui, Xi Wang, Lei He, and Frank K. Soong

Microsoft China
{yancu,xwang,helei,frankkps}@microsoft.com

Abstract

LPCNet neural vocoder and its variants have shown the ability to synthesize high-quality speech in small footprint by exploiting domain knowledge in speech. In this paper, we introduce subband linear prediction in LPCNet for producing high fidelity speech more efficiently with consideration of subband correlation. Speech is decomposed into multiple subband signals with linear prediction to reduce the complexity of neural vocoder. A novel subband-based autoregressive model is proposed to learn the joint distribution of the subband sequences by introducing a reasonable assumption, which keeps the dependence between subbands while accelerating the inference speed. Based upon the human auditory perception sensitivity to the harmonic speech components in the baseband, we allocate more computational resources to model the low-frequency subband to synthesize natural phase and magnitude of the synthesized speech. Both objective and subjective tests show the proposed subband LPCNet neural vocoder can synthesize higher quality speech than the original fullband one (MOS 4.62 vs. 4.54), at a rate nearly three times faster.

Index Terms: subband LPCNet, neural vocoder, speech synthesis, subband linear prediction, multirate signal processing

1. Introduction

Deep generative models have been successfully applied to the text-to-speech system and improved the quality of synthesized speech towards human parity. The typical neural vocoders, such as WaveNet[1], WaveRNN[2] , and WaveGlow[3], can produce high-fidelity speech by using kinds of generative neural models. However, it is still a challenging task for real-time synthesis on the servers and devices with a typical configuration.

WaveNet can deliver natural and high-fidelity speech with the autoregressive generative model. Nonetheless, the inference speed is much slower than real-time synthesis due to the sequential generation mechanism. Parallel WaveNet[4] and the ClariNet[5] make the inference in parallel through the distillation from a pre-trained teacher WaveNet. But it is difficult to implement and deploy due to the complicated training pipeline. Besides, it requires the GPU with high computational performance for real-time synthesis because the total computational complexity of generation is not reduced. WaveGlow directly optimizes the likelihood of latent variables with bipartite flows, which is simple to implement and train. But it requires much more parameters than WaveNet to reach comparable quality. WaveRNN improves the inference efficiency of the recurrent network by a series of techniques of compression and generation. Moreover, it also shows the robustness in several scenarios[6, 7]. But the synthesis quality degrades as the model size decreases[8].

Some recent works have improved the performance of neural vocoder by introducing linear prediction (LP) analysis. Linear prediction analysis decouples the speech into the spectral envelope and the excitation signal, which makes it simpler to predict the excitation distribution instead of speech. Based on the source-filter model, some neural vocoders can get natural speech with smaller model and higher efficiency, such as LPCNet[8], LP-WaveNet[9], neural source-filter model[10], and glottal neural vocoders[11, 12, 13, 14]. Specifically, the LPCNet vocoder achieves higher quality than WaveRNN for quite small model size. An improvement of the LPCNet, the iLPCNet[15] presents a closed-loop solution for the excitation and spectral modeling and further improves the synthesis quality with 16-bit linear PCM.

The multirate signal processing can also make a significant contribution to improve the inference efficiency by decomposing the speech into the subband domain. The advantages of performing speech generation in the subband domain include: 1. The length of each subband is much shorter than fullband speech by down sampling, which reduces the complexity and improves the efficiency. 2. Subband decomposition facilitates the manipulation of each subband based upon the human auditory perception sensitivity. Some related works have investigated the neural vocoders in the subband domain, such as subband WaveNet[16, 17, 18], subband FFTNet[19], subband WaveRNN[20], FeatherWave[21], and multiband MelGan[22]. To produce the subband sequences simultaneously, most of these works take the independence of each subband signal as a default assumption. However, such an assumption will lead to the mismatch between training and inference, which affects the final quality of output speech.

To further improve the synthesis efficiency and quality, we propose the subband LPCNet, a high-fidelity neural vocoder by the combination of series of speech domain knowledge, e.g. multirate signal processing, linear prediction analysis, and human auditory perception. First, the speech signal is decomposed into four subband signals through the analysis filter banks. A novel subband-based autoregressive model is proposed to model the joint distribution of the subband sequences with consideration of the subband correlation. Furthermore, the excitation signal is extracted by leveraging linear prediction analysis in the baseband, which makes a significant contribution to the human auditory perception of synthesized speech. Inspired from speech coding[23], we allocate more computational resources to the baseband module of the neural vocoder. The proposed neural vocoder takes the acoustic feature as condition and generates the subband samples simultaneously without the mismatch between training and inference. Finally, the fullband synthesized speech is reconstructed from the generated subband samples through the synthesis filter banks.

This paper is organized as follows. Section 2 introduces a novel subband-based autoregressive model for speech generation. Section 3 presents the structure of the proposed neural vocoder, which is followed by the evaluation results in Section 4. Finally, we draw the conclusions in Section 5.

http://dx.doi.org/10.21437/Interspeech.2020-1463

2. Subband-based autoregressive model

The WaveNet vocoder models the probability distribution of the speech sequence $p(\boldsymbol{x};\boldsymbol{h})$ conditioned on the product of the history sequence distribution, given acoustic features $\boldsymbol{h}$, as:

$$p(\boldsymbol{x};\boldsymbol{h}) = \prod_{t=1}^{T} p(x_t|\boldsymbol{x}_{<t};\boldsymbol{h}). \tag{1}$$

In a multirate system, the speech is decomposed into several subband streams through analysis filter banks and reconstructed by synthesis filter banks. The autoregressive model of time domain can also be expended in the subband domain, as:

$$p(X;\boldsymbol{h}) = \prod_{k=1}^{K} p(x_{1,k},\, x_{2,k},\, x_{3,k},\, x_{4,k}|X_{\bullet,<k};\boldsymbol{h}), \tag{2}$$

where $X \in \mathbb{R}^{D\times K}$ represents the matrix composed of the subband sequences. D is the decimate rate and K is the sequence length of each subband. Without loss of generality, let $D = 4$ here. $X_{\bullet,<k} = [\boldsymbol{x}_{1,<k}, \boldsymbol{x}_{2,<k}, \boldsymbol{x}_{3,<k}, \boldsymbol{x}_{4,<k}]$ represents the submatrix of X before the k^{th} column. $x_{i,k}$ denotes the k^{th} sample of the i^{th} subband sequence. $\boldsymbol{x}_{i,<k}$ denotes all the samples before the k^{th} one in the i^{th} subband sequence. The first subband is the lowest frequency band.

Previous works take the sampling of each subband distribution separately during inference, which takes the independence of all the subband distributions as a default assumption. However, all of the subband signals are highly correlated with each other[24]. The independent assumption will lead to a phase mismatch on the boundary between the neighbor subbands[16].

In this paper, we address that the joint distribution of the subband sequence is conditioned on the lower frequency band. By introducing such an assumption, the joint distribution $p(x_{1,k}, x_{2,k}, x_{3,k}, x_{4,k})$ can be decomposed as:

$$\begin{aligned} p(x_{1,k},\, x_{2,k},\, x_{3,k},\, x_{4,k}) = p(x_{1,k})p(x_{2,k}|x_{1,k}) \\ p(x_{3,k}|x_{1,k},\, x_{2,k})p(x_{4,k}|x_{1,k},\, x_{2,k},\, x_{3,k}). \end{aligned} \tag{3}$$

Let $\boldsymbol{y}_{1,k}$, $\boldsymbol{y}_{2,k-1}$, $\boldsymbol{y}_{3,k-2}$, and $\boldsymbol{y}_{4,k-3}$ denote the history conditional samples of corresponding subband, respectively:

$$\begin{aligned} \boldsymbol{y}_{1,k} &= \boldsymbol{x}_{1,<k}, \\ \boldsymbol{y}_{2,k-1} &= [x_{1,k-1},\, \boldsymbol{x}_{2,<k-1}], \\ \boldsymbol{y}_{3,k-2} &= [x_{1,k-2},\, x_{2,<k-2},\, \boldsymbol{x}_{3,<k-2}], \\ \boldsymbol{y}_{4,k-3} &= [x_{1,k-3},\, x_{2,k-3},\, x_{3,k-3},\, \boldsymbol{x}_{4,<k-3}]. \end{aligned} \tag{4}$$

Substitute Eq.(3) into Eq.(2), the subband autoregressive model is:

$$\begin{aligned} p(X;\boldsymbol{h}) = \prod_{k=1}^{K+3} p(x_{1,k}|\boldsymbol{y}_{1,k};\boldsymbol{h})p(x_{2,k-1}|\boldsymbol{y}_{2,k-1};\boldsymbol{h}) \\ p(x_{3,k-2}|\boldsymbol{y}_{3,k-2};\boldsymbol{h})p(x_{4,k-3}|\boldsymbol{y}_{4,k-3};\boldsymbol{h}). \end{aligned} \tag{5}$$

Fig. 1 shows the corresponding probabilistic graph of the subband sequences. The probability distribution of subband samples $p(x_{1,k})$, $p(x_{2,k-1})$, $p(x_{3,k-2})$ and $p(x_{4,k-3})$ (the light green nodes) are coherently independent given $\boldsymbol{x}_{1,<k}$, $\boldsymbol{x}_{2,<k-1}$, $\boldsymbol{x}_{3,<k-2}$, and $\boldsymbol{x}_{4,<k-3}$ (the dark blue nodes). So that the predicted distribution of each subband signal can be sampled separately to improve both the inference efficiency and precision without the mismatch. Notice that the subband generative model predicts the subband sequence by a few samples delay. It is necessary to compensate for the corresponding delay in each subband before the reconstruction through synthesis filter banks.

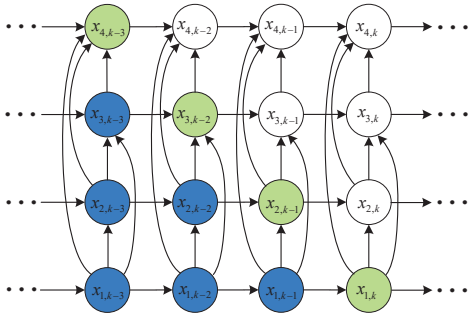
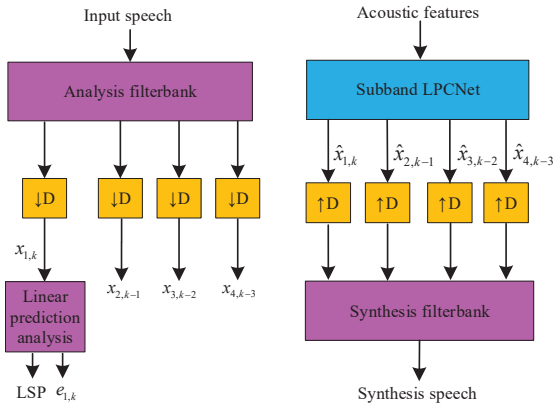


Figure 1: *The probabilistic graph of subband sequences. Each node corresponds one subband sample. The arrow lines indicate the dependency among the sequences. The light green nodes are independent given the dark blue nodes.*



(a) Analysis stage (b) Synthesis stage

Figure 2: *Block diagram of analysis and synthesis stage of signal processing.*

3. Proposed neural vocoder

The proposed neural vocoder consists of two modules: the signal processing and the neural network modules. The signal processing module can reduce the complexity of neural module by signal processing, including acoustic feature extraction, multirate signal processing and linear prediction analysis. The neural network module learns the distribution of subband samples conditioned on the acoustic features and generate the subband sequence autoregressively.

3.1. Signal processing module

Fig. 2 shows the block diagram of analysis and synthesis. Firstly, the Bark-scale frequency cepstral coefficients (BFCC), Log-scale F0 and pitch correlation are extracted from the full-band speech as acoustic features. An efficient method is introduced to design the analysis and synthesis filter banks for multirate signal processing[25]. In the training stage, the speech is decomposed into four equal bandwidth subbands in linear frequency through the analysis filter banks. The subband sequences are rearranged according to Eq.(5). Then linear prediction analysis is leveraged to extract the linear spectral pair (LSP) features and the excitation signals in the first subband.

In synthesis, the subband distribution is predicted by the neural module conditioned on the acoustic features. After autoregressive sampling, the first subband excitation and other subband samples are generated. The LP coefficients of the first subband is calculated from the LSP features. Then the first subband samples are generated by passing the excitation signal through the LP synthesis filter. Finally, the fullband synthesized speech is reconstructed through the synthesis filter banks. The linear prediction in the first subband is as follows:

$$\begin{aligned} x_{1,k} &= e_{1,k} + p_{1,k}, \\ p_{1,k} &= \sum_{i=1}^{M} a_i \, x_{1,k-i}, \end{aligned} \tag{6}$$

where $x_{1,k}$, $e_{1,k}$, and $p_{1,k}$ denote the subband signal, the excitation signal and the linear prediction of the first subband, respectively. a_i denotes the i^{th} LP coefficients with the order M.

3.2. Neural network module

The neural network module consists of two networks at two rates, i.e., frame and sample rates, as Fig. 3 shows. The frame rate network processes the acoustic features as condition and upsamples them then sends the results to the sample rate network. Firstly, F0 is linearly quantized to 256 discrete levels, passed through an embedding layer then concatenated with the acoustic features. A two-layer convolutional neural network is adopted to construct the context information across the adjacent five frames. After one residual connection and one fully connected (FC) layer, the three FC layers transform the context vector as the input of three recurrent units in sample rate network. We expect these three conditions will focus on different necessary information to model corresponding subbands.

The sample rate network takes the three outputs of the frame rate network as condition and models the distribution of all subband samples. All the subband samples are quantized to 8-bit with μ-law compression and share one embedding layer. Because they are all waveforms in the same sampling rate with the same physical meaning. Conditioned by the frame rate network, a gated recurrent unit (GRUa) is shared by four subband samples at previous timestep $\boldsymbol{x}_{\cdot,k-1} = [x_{1,k-1}, x_{2,k-2}, x_{3,k-3}, x_{4,k-4}]$. After GRUa, two smaller recurrent units (GRUb and GRUc) learn the distribution of the first subband and the other three subbands. GRUb takes the excitation sample at previous timestep $e_{1,k-1}$ as an additional condition. Based upon the human auditory perception sensitivity of the baseband, the first subband contributes more to the phase consistance and formant shape than other three subbands. So we pay more attention to the first subband quality and allocate more computational resources to it. To further improve the precision, logistic mixture density network (MDN) is utilized to learn the distribution of the excitation signal in the first subband. Besides, MDN with linear prediction presents a closed-loop solution to model the excitation and the subband samples in the first subband[15]. The loss function is the non-negative log-likelyhood (NLL) for the first subband and the cross entropy (CE) for the other three subbands. The MDN parameters and

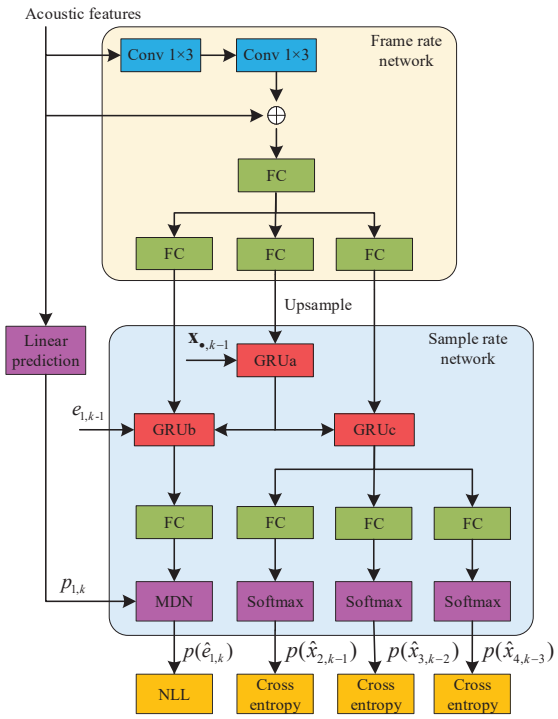


Figure 3: *Block diagram of the neural network module of proposed neural vocoder.*

the loss function are shown as follows:

$$\begin{aligned} p(\hat{e}_{1,k}) &= \sum_{i=1}^{N} \omega_{k,i}^{e} \text{logistic}(\mu_{k,i}^{e}, \, s_{k,i}^{e}), \\ \mathcal{L} &= \mathcal{L}_{\text{nll}} + \lambda \mathcal{L}_{\text{ce}}, \\ \mathcal{L}_{\text{nll}} &= -\log p(\hat{e}_{1,k}), \\ \mathcal{L}_{\text{ce}} &= -\sum_{i=2}^{4} p(x_{i,k-i+1}) \log p(\hat{x}_{i,k-i+1}), \end{aligned} \tag{7}$$

where $\omega_{k,i}^{e}$, $\mu_{k,i}^{e}$ and $s_{k,i}^{e}$ denote the i^{th} mixture parmeters (proportion, mean and variance) of the first subband excitation signal, respectively. N is the mixture number. $\text{logistic}()$ denotes the logistic distribution function. $\mathcal{L}$ is the loss function. $\mathcal{L}_{\text{nll}}$ is the NLL loss and $\mathcal{L}_{\text{ce}}$ is the CE loss. λ is a factor that controls the balance between the first subband and the other three subbands.

In synthesis, the neural network module takes the acoustic features as the condition and predicts the subband samples at each timestep autoregressively. Specifically, the first subband sample is generated by adding the excitation to the linear prediction as Eq.(6).

4. Experiments

We perform the experiments on the analysis-by-synthesis samples, and evaluate them by both the objective and subjective tests. The computational cost of the proposed subband LPCNet is also calculated and compared with the original fullband LPCNet vocoder at the sampling rate of 24kHz.

Table 1: *Acoustic features and dimension.*

Feature name	Dimension
BFCC	20
Log-scale F0	1
Pitch correlation	1
LSP	8
LPC	8

4.1. Database and setup

The experiments are performed on a database which consists of over 11,000 sentences for approximately 12 hours recordings by a professional US female speaker. We chose 10,000, 500, and 500 utterances as training, validation, and test sets, respectively. The speech signals are sampled at 24kHz with 16-bit linear PCM.

A pre-emphasis filter is applied before the analysis filter banks and the fullband speech is reconstructed with a matched de-emphasis filter after the synthesis filter banks. The acoustic features for the neural module are BFCC, F0, pitch correlation and LSP feautures. The first three acoustic features are extracted from the fullband speech, and the LSP features and corresponding LPC features are extracted from the first subband signal. The dimension of each acoustic feature is listed in Table 1. The frameshift of acoustic features is 10ms (240 samples). The noise injection and the sparsity of recurrent units are as same as the original fullband LPCNet setup[8].

4.2. Neural module and the complexity

The dimension of GRUa, GRUb, and GRUc in the sample rate network are 384, 16, and 16, respectively. The dimension of FC layers is set to 128. The convolutional layers in the frame rate network have 1×3 kernels and 128 channels. The weighting factor λ in the loss function is set to 0.5. In training stage, the proposed neural vocoder is trained on a GPU cluster with 4 GPUs. The Adam optimizer is adopted and initialized with a learning rate of 0.001. The mini-batch size is 32 and the training iteration is 120 epochs.

The number of weights in the framework rate network W_f and sample rate network W_s are:

$$\begin{aligned} W_f &= 2N_{conv}N_{ac} + 4N_{fc}^2, \\ W_s &= 3(dN_a^2 + N_a(N_b + N_c) + N_bL + N_cQ), \end{aligned} \tag{8}$$

where $N_a = 384$, $N_b = 16$, and $N_c = 16$ denote the size of GRUa, GRUb, and GRUc, respectively. $d = 0.1$ is the density of the sparse GRU, $Q = 256$ is the quantization level, and $L = 10$ is the mixture number of MDN. $N_{conv} = 128$ and $N_{fc} = 128$ are the dimension of the convolutional layer and the FC layer. $N_{ac} = 30$ is the dimension of the acoustic feauture.

We leverage float-point operations per second (FLOPS) as a measurement of the total complexity of neural vocoders, which shows the float-point operations required to synthesize one second speech waveform. The total complexity of the proposed subband LPCNet is about 1.2 GFLOPS. And the complexity of the original fullband LPCNet is around 3.5 GFLOPS. As a comparison, the total complexity of fullband WaveRNN and WaveNet vocoder is 10 GFLOPS and 100 GFLOPS[26] approximately. The proposed subband LPCNet is nearly three times faster than the original fullband LPCNet, which makes it practical and effective to deploy.

Table 2: *The results of the objective and subjective tests.*

Voice name	LSD(dB)	MCD(dB)	MOS
Recording	–	–	4.73 ± 0.05
Subband LPCNet(w/)	7.52	3.03	4.62 ± 0.05
Subband LPCNet(w/o)	7.56	3.21	4.60 ± 0.05
Fullband LPCNet	7.85	4.03	4.54 ± 0.05

4.3. Evaluations

The original fullband LPCNet vocoder is chosen to compare with the proposed neural vocoder (subband LPCNet w/). Besides, a neural vocoder with the same structure as the proposed vocoder (subband LPCNet w/o) is also included in the experiments as a comparison. The only difference is that the subband LPCNet vocoder leverages the proposed subband-based generative model and the other one still takes the independent assumption among all the subbands. All test samples of synthesized speech are generated through the analysis-by-synthesis process. Some test samples are available at the following link[1].

To objetively measure the performance of the proposed neural vocoder, we computed both the logarithm spectral distance (LSD) and the Mel-cepstral distortion (MCD) between the natural recordings and the test samples. The results of objective tests are shown in Table 2. It can be observed that the proposed subband LPCNet(w/) outperforms the other two vocoders in the measurement of LSD and MCD.

We also evaluated the perceptual quality of the proposed neural vocoder by a subjective mean opinion score (MOS) test. In the test, 50 utterances are randomly selected from the evaluation set and 20 listeners are asked to give a natural score from 1 (Bad) to 5 (Excellent) of a perceived sample. The MOS test result in Table 2 shows that the proposed subband LPCNet with the subband-based autoregressive model gets the highest score in the three tested vocoders. The neural vocoder with the independent assumption is slightly worse than the one with the proposed subband LPCNet, which is consistent with the objective evaluation results. These experimental results could be attributed to the subband decomposition and the well-designed structure, which decomposes the fullband speech into several subbands with much shorter sequence length and reduces the modeling complexity.

5. Conclusions

In this paper, we propose the subband LPCNet, a highly efficient neural vocoder with subband-based linear prediction for speech synthesis. Its computational complexity is significantly reduced by exploiting domain knowledge of speech signals in multirate signal processing, linear prediction analysis, and human auditory perception in different frequency bands. To alleviate the mismatch between the training and inference, a novel subband autoregressive model is proposed to improve the inference efficiency while maintaining the synthesized speech quality. Both objective and subjective experiments show that the proposed subband LPCNet vocoder outperforms the original fullband LPCNet vocoder while accelerating the inference speed at a rate three times faster.

In the future, we will continue to investigate the performance of the subband LPCNet with an end-to-end, text-to-speech synthesis system and introduce more knowledge of speech coding to improve the efficiency and quality of synthesized speech.

[1] https://scimagian.github.io/subbandLPCNet/demo/index.html

6. References

[1] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[2] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. v. d. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," *arXiv preprint arXiv:1802.08435*, 2018.

[3] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3617–3621.

[4] A. v. d. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. v. d. Driessche, E. Lockhart, L. C. Cobo, F. Stimberg *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," *arXiv preprint arXiv:1711.10433*, 2017.

[5] W. Ping, K. Peng, and J. Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," *arXiv preprint arXiv:1807.07281*, 2018.

[6] J. Lorenzo-Trueba, T. Drugman, J. Latorre, T. Merritt, B. Putrycz, R. Barra-Chicote, A. Moinet, and V. Aggarwal, "Towards achieving robust universal neural vocoding," *arXiv preprint arXiv:1811.06292*, 2018.

[7] P.-c. Hsu, C.-h. Wang, A. T. Liu, and H.-y. Lee, "Towards robust neural vocoding for speech generation: A survey," *arXiv preprint arXiv:1912.02461*, 2019.

[8] 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.

[9] M.-J. Hwang, F. Soong, F. Xie, X. Wang, and H.-G. Kang, "Lp-wavenet: Linear prediction-based wavenet speech synthesis," *arXiv preprint arXiv:1811.11913*, 2018.

[10] X. Wang, S. Takaki, and J. Yamagishi, "Neural source-filter-based waveform model for statistical parametric speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5916–5920.

[11] M. Airaksinen, B. Bollepalli, L. Juvela, Z. Wu, S. King, and P. Alku, "Glottdnn-a full-band glottal vocoder for statistical parametric speech synthesis." in *Interspeech*, vol. 9, 2016, pp. 2473–2477.

[12] Y. Cui, X. Wang, L. He, and F. K. Soong, "A new glottal neural vocoder for speech synthesis." in *Interspeech*, 2018, pp. 2017–2021.

[13] L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, "Gelp: Gan-excited liner prediction for speech synthesis from mel-spectrogram," *arXiv preprint arXiv:1904.03976*, 2019.

[14] L. Juvela, B. Bollepalli, V. Tsiaras, and P. Alku, "Glotnet—a raw waveform model for the glottal excitation in statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 1019–1030, 2019.

[15] M.-J. Hwang, E. Song, R. Yamamoto, F. Soong, and H.-G. Kang, "Improving lpcnet-based text-to-speech with linear prediction-structured mixture density network," *arXiv preprint arXiv:2001.11686*, 2020.

[16] T. Okamoto, K. Tachibana, T. Toda, Y. Shiga, and H. Kawai, "Subband wavenet with overlapped single-sideband filterbanks," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 698–704.

[17] ——, "An investigation of subband wavenet vocoder covering entire audible frequency range with limited acoustic features," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5654–5658.

[18] A. Rabiee, G. Kim, T.-H. Kim, and S.-Y. Lee, "A fully time-domain neural model for subband-based speech synthesizer," *arXiv preprint arXiv:1810.05319*, 2018.

[19] T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, "Improving fftnet vocoder with noise shaping and subband approaches," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 304–311.

[20] C. Yu, H. Lu, N. Hu, M. Yu, C. Weng, K. Xu, P. Liu, D. Tuo, S. Kang, G. Lei *et al.*, "Durian: Duration informed attention network for multimodal synthesis," *arXiv preprint arXiv:1909.01700*, 2019.

[21] Q. Tian, Z. Zhang, L. Heng, L. Chen, and S. Liu, "Feather wave: An efficient high-fidelity neural vocoder with multiband linear prediction," *arXiv preprint arXiv:arXiv:2005.05551*, 2020.

[22] G. Yang, S. Yang, K. Liu, P. Fang, W. Chen, and L. Xie, "Multi-band melgan: Faster waveform generation for high-quality text-to-speech," *arXiv preprint arXiv:arXiv:2005.05106*, 2020.

[23] K. Brandenburg and G. Stoll, "Iso/mpeg-1 audio: A generic standard for coding of high-quality digital audio," *Journal of the Audio Engineering Society*, vol. 42, no. 10, pp. 780–792, 1994.

[24] J. McAuley, J. Ming, D. Stewart, and P. Hanna, "Subband correlation and robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 956–964, 2005.

[25] C. D. Creusere and S. K. Mitra, "A simple method for designing high-quality prototype filters for m-band pseudo qmf banks," *IEEE Transactions on signal processing*, vol. 43, no. 4, pp. 1005–1007, 1995.

[26] J.-M. Valin and J. Skoglund, "A real-time wideband neural vocoder at 1.6 kb/s using lpcnet," *arXiv preprint arXiv:1903.12087*, 2019.