HIGH-QUALITY SPEECH CODING WITH SAMPLE RNN

Janusz Klejsa [*]   *Per Hedelin* [*]   *Cong Zhou* [†]   *Roy Fejgin* [†]   *Lars Villemoes* [*]

[*] Dolby Sweden AB, Stockholm, Sweden
[†] Dolby Laboratories, San Francisco, CA, USA

ABSTRACT

We provide a speech coding scheme employing a generative model based on SampleRNN that, while operating at significantly lower bitrates, matches or surpasses the perceptual quality of state-of-the-art classic wide-band codecs. Moreover, it is demonstrated that the proposed scheme can provide a meaningful rate-distortion trade-off without retraining. We evaluate the proposed scheme in a series of listening tests and discuss limitations of the approach.

***Index Terms*—** speech coding, deep neural networks, vocoder, SampleRNN

1. INTRODUCTION

Generative modeling for audio based on deep neural networks, such as WaveNet [1] and SampleRNN [2], has provided significant advances in natural-sounding speech synthesis. The main application has been in the field of text-to-speech [3, 4, 5] and parametric synthesis [6], where the models replace the vocoding component.

Moreover, generative models can be conditioned by global and local latent representations [1]. In the context of voice conversion, this facilitates natural separation of the conditioning into a static speaker identifier and dynamic linguistic information [7].

In this paper we develop a wide-band conditioning from the vocoder [8] with parameters quantized to 5.6, 6.4 and 8 kb/s and use these to condition a SampleRNN model. We benchmark our resulting speech coding scheme against AMR-WB [9] at 23.05 kb/s and SILK [10] at 16 kb/s, which, at this bitrate, is a state-of-the-art classic speech codec providing good to excellent quality [11].

While it was shown in [12] that the subjective quality of AMR-WB at 23.05 kb/s can be approached at only 2.4 kb/s by using WaveNet synthesis conditioned on narrowband vocoder parameters, the quality gap to classic speech coding schemes was not closed. Furthermore, the reconstructed waveforms occasionally suffered from typical blind bandwidth extension artifacts (the WaveNet-based decoder used narrow-band conditioning but synthesized wide-band speech).

The purpose of this paper is two-fold. First, we demonstrate that the approach of [12] is reproducible with another generative network architecture. Second, we investigate the scalability of such a scheme in terms of rate-quality trade-off. In particular, we demonstrate that higher reconstruction quality can be achieved by allowing a higher bitrate for the conditioning. The subjective quality is evaluated with a methodology based on MUSHRA[13].

The speech coder structure considered in this paper is shown in Fig. 1 at a high level. It includes an encoder based on a vocoding structure governed by a parametric signal model that facilitates quantization constrained by bitrate. The decoder scheme uses a four-tier SampleRNN architecture. The conditioning parameters are designed in a way that lower quality conditioning can be embedded into higher quality conditioning, which allows for fixed dimensionality of the conditioning irrespectively of the operating bitrate. The signal sample modeling and synthesis are based on a discretized logistic mixture [14] instead of the 8-bit μ-law domain used in [1, 12].

Perhaps one of the most interesting questions related to the proposed scheme is whether its performance generalizes to unseen speech signals. While it is relatively straightforward to avoid model overfitting within the selected dataset by using a state-of-the-art training approach (e.g., [15, 16]) the issue of robustness of the coding algorithm remains unclear. In order to facilitate comparison to [12] we carried our experiments with the same multi-speaker dataset, namely the Wall Street Journal set WSJ0 [17]. However, it comprises only American English, and seems to have been created with capture devices of limited quality. Hence, we cross-validated the performance of the proposed approach on another publicly available dataset (clean speech from the VCTK corpus [18]). We demonstrate how the performance of the coding scheme degrades on signals from the new data set and then improves again with retraining on an extended dataset.

This paper is structured as follows. First, the vocoder and its quantization scheme are described in Section 2 along with the embedding approach. Next, we describe our SampleRNN model and provide the details of its vocoder conditioning in Section 3. The experimental setup and the listening test results are presented in Section 4. Finally, conclusions are presented in Section 5.

2. VOCODER WITH QUANTIZED PARAMETERS

The encoder scheme is based on a wide-band version of a linear prediction coding (LPC) vocoder [8]. Signal analysis is performed on a per-frame basis, and it results in the following parameters: *i)* an M-th order LPC filter, *ii)* an LPC residual RMS level s, *iii)* pitch f_0, and, *iv)* a k-band voicing vector v. A voicing component $v(i)$, $i = 1, \ldots, k$ gives the fraction of periodic energy within a band. All these parameters are used for conditioning of SampleRNN, as described in Section 3. We note the signal model used by the encoder aims at describing only clean speech (without background or
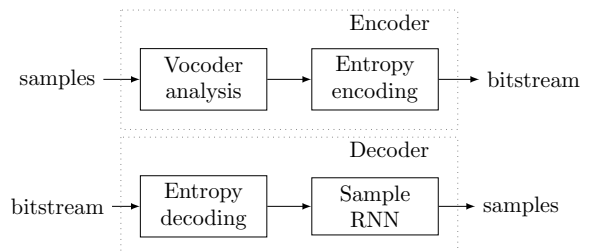

Fig. 1. Block diagrams of the vocoder-based encoder and the SampleRNN decoder.

Table 1. Operating points of the encoder ($k = 6$)

r_{nominal} [kb/s]	M	spectral dist. [dB]	n bits s	n bits v_w
8.0	22	0.754	$1 + 9$	9
6.4	16	0.782	$1 + 8$	9
5.6	16	1.33	$1 + 8$	9

simultaneously active talkers).

The analysis scheme operates on 10 ms frames of a signal sampled at 16 kHz. In the proposed encoder design, the order of the LPC model, M, depends on the operating bitrate. Standard combinations of source coding techniques are utilized to achieve encoding efficiency with appropriate perceptual consideration, including vector quantization (VQ), predictive coding and entropy coding [19]. In this paper, for all experiments, we define the operating points of the encoder as in Table 1. We used standard tuning practices. For example, the spectral distortion for the reconstructed LPC coefficients is kept close to 1 dB [20].

The LPC model is coded in the line spectral pairs (LSP) domain utilizing prediction and entropy coding. For each LPC order, M, a Gaussian mixture model (GMM) was trained on the WSJ0 train set, providing probabilities for the quantization cells. Each GMM component has a $\mathcal{Z}$-lattice according to the principle of union of $\mathcal{Z}$-lattices [21, 22]. The final choice of quantization cell is according to a rate-distortion weighted criterion.

The residual level s is quantized in the dB domain using a hybrid approach similar to that in [23]. Small level inter-frame variations are detected, signalled by one bit, and coded by a predictive scheme using fine uniform quantization. In other cases the coding is memoryless with a larger, yet uniform, step-size covering a wide range of levels.

Similar to level, pitch is quantized using a hybrid approach of predictive and memoryless coding. Uniform quantization is employed but executed in a warped pitch domain. Pitch is warped by $f_w = c f_0 / (c + f_0)$ where $c = 500$ Hz and f_w is quantized and coded using 10 bit/frame.

Voicing is coded by memoryless VQ in a warped domain. Each voicing component is warped by $v_w(i) = \log(\frac{1 - v(i)}{1 + v(i)})$. A 9 bit VQ was trained in the warped domain on the WSJ0 train set.

A feature vector $\mathbf{h_f}$ for conditioning SampleRNN is constructed as follows. The quantized LPC coefficients are converted to reflection coefficients. The vector of refection coefficients is concatenated with the other quantized parameters, i.e. f_0, s, and v. In the remainder of the paper we use either of two constructions of the conditioning vector. The first construction is the straightforward concatenation described above. For example, for $M = 16$, the total dimension of the vector $\mathbf{h_f}$ is 24; for $M = 22$ it is 30. The second construction is an embedding of lower-rate conditioning into a higher-rate format. For example, for $M = 16$, an 22-dimensional vector of the reflection coefficients is constructed by padding the 16 coefficients with 6 zeros. The remaining parameters are replaced with their coarsely quantized (low bitrate) versions, which is possible since their locations within $\mathbf{h_f}$ are now fixed.

3. CONDITIONAL SAMPLE RNN

SampleRNN is a deep neural generative model proposed in [2] for generating raw audio signals. It consists of a series of multi-rate recurrent layers, which are capable of modeling the dynamics of a sequence at different time scales. SampleRNN models the proba-
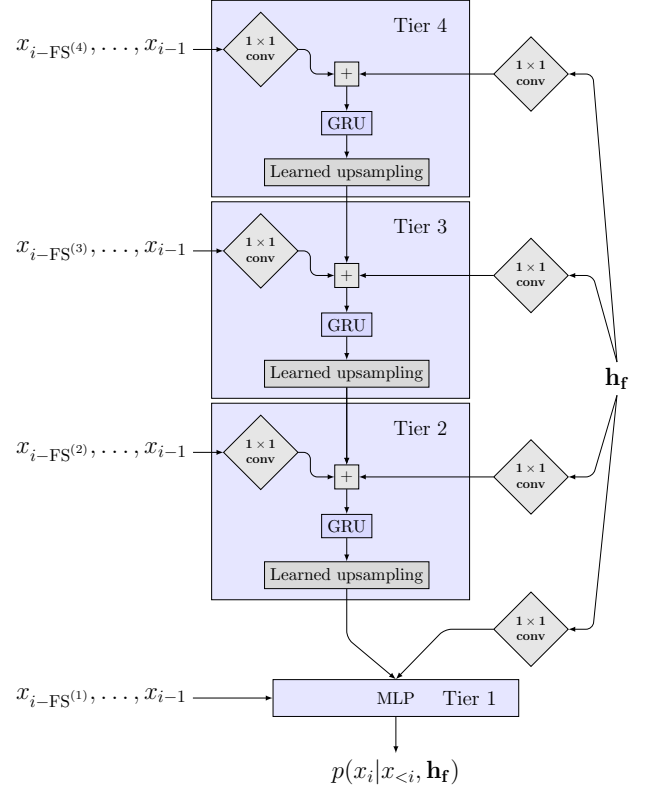
Fig. 2. Conditioning of the SampleRNN.

bility of a sequence of audio samples via factorization of the joint distribution into the product of the individual audio sample distributions conditioned on all previous samples. The joint probability distribution of a sequence of waveform samples $X = \{x_1, \ldots, x_T\}$ can be written as

$$p(X) = \prod_{i=1}^{T} p(x_i | x_1, \ldots, x_{i-1}). \tag{1}$$

At inference time, the model predicts one sample at a time by randomly sampling from $p(x_i | x_1, \ldots, x_{i-1})$. Recursive conditioning is then performed using the previously reconstructed samples.

3.1. Conditioning

Without conditioning information, SampleRNN is only capable of "babbling". Hence, we provide decoded vocoder parameters, $\mathbf{h_f}$, as conditioning information to the model. Eq. 1 thus becomes

$$p(X|H) = \prod_{i=1}^{T} p(x_i | x_1, \ldots, x_{i-1}, \mathbf{h_f}), \tag{2}$$

where $\mathbf{h_f}$ represents the vocoder parameters corresponding to the audio sample at time i.

The structure of a conditional SampleRNN is illustrated in Fig. 2. In a K-tier conditional SampleRNN, the k-th tier ($1 < k \leq K$) operates on non-overlapping frames of length FS$^{(k)}$ samples at a time, and the lowest tier ($k = 1$) predicts one sample at a time. Waveform samples $x_{i-\text{FS}^{(k)}}, \ldots, x_{i-1}$ and decoded vocoder conditioning vector $\mathbf{h_f}$ processed by respective 1×1 convolution layers
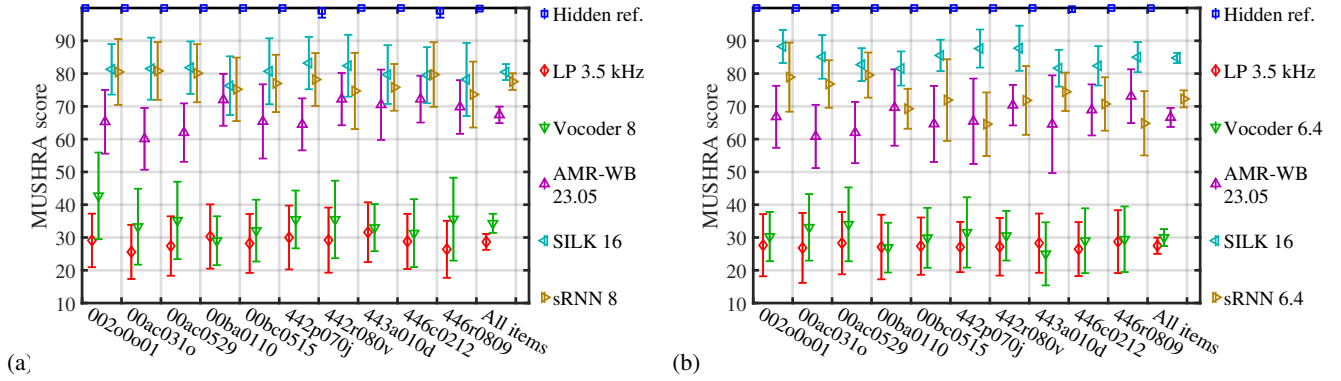
Fig. 3. Results of the listening test comparing sRNN and vocoder (operating at 8 kb/s (a) and 6.4 kb/s (b)) and the reference speech codecs: AMR-WB (23.05 kb/s) and SILK (16 kb/s); there were 10 listeners in the test (a), and 11 in the test (b); we use original names of the excerpts.

are the inputs to k-th tier. When $k < K$, the output from $(k + 1)$-th tier is additional input. Similar to [7], all inputs to k-th tier are linearly summed up. The k-th RNN tier $(1 < k \leq K)$ consists of one gated recurrent unit (GRU) layer, cf. [24], and one learned up-sampling layer performing temporal resolution alignment between tiers. The lowest $(k = 1)$ tier consists of a multilayer perceptron (MLP) with 2 hidden fully connected layers.

3.2. Discretized logistic mixture

Instead of using 256-way softmax in the 8-bit μ-law domain as in [5, 12], we adopt the discretized mixture of logistic distributions technique introduced in [14, 25] to generate 16-bit samples. The output layer of the sample-level MLP predicts parameters of each mixture component, including mixture weight, mean, and log scale. Unlike the original SampleRNN, which performs an embedding of quantized sample values into real-valued vectors, we use the raw samples directly as the input to the MLP tier.

3.3. Model configurations and training setup

As shown in Fig. 2, our architecture has a four-tier configuration $(K = 4)$, where the frame size for the k-th tier is $FS^{(k)}$. We use the following frame sizes: $FS^{(1)} = FS^{(2)} = 2$, $FS^{(3)} = 16$ and $FS^{(4)} = 160$. The top tier shares the same temporal resolution as the vocoder parameter conditioning sequence. The learned upsampling layer is implemented through a transposed convolution layer, and the upsampling ratio is 2, 8, and 10 respectively in the second, third and fourth tier. The recurrent layers and fully connected layers contain 1024 hidden units each.

During training, we use a batch size of 24 and a sequence length of 6400 samples for truncated back propagation through time (TBPTT, [26, 27]) on a single GPU. We use the ADAM optimizer [28] ($\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = $ 1e-8) with an initial learning rate of 0.0002. The learning rate is reduced by multiplying by a factor 0.3 when validation loss has stopped dropping. Gradients are hard-clipped to a range of $[-1, 1]$.

4. EXPERIMENTAL EVALUATION

4.1. Perceptual evaluation on the WSJ0 set

We performed perceptual evaluation of the our SampleRNN-based decoder on randomly selected excerpts from the test set of WSJ0 in a listening test resembling the experiment in [12]. There were 5 male

speakers and 5 female speakers. The items in the original WSJ0 data set have inconsistent loudness, which is problematic for a listening test. Thus, we performed loudness alignment of the test items prior to encoding.

We used a headphone based MUSHRA-like test methodology [13]. Since the speech data set was only available at 16 kHz sampling frequency, we only included the 3.5 kHz low-pass anchor (LP 3.5 kHz), a wide-band original and hidden reference, and no 7 kHz anchor. The conditions used in the test included SILK (at 16 kb/s, variable rate), which is a state-of-the-art speech codec, and AMR-WB (at 23.05 kb/s, constant bitrate) in a configuration that facilitates comparison of our results to the 2.4 kb/s decoder of [12]. In the first two listening tests, we included pairs of conditions: reconstruction provided by the SampleRNN (sRNN) decoder, and reconstuction provided by the vocoder [8], which was used to produce the conditioning. Thus both conditions represent decoding of the same bitstream. In the two following experiments, we used sRNN decoders that were trained specifically for their respective operating bitrates.

In the first experiment, we performed a test where the vocoder and the sRNN decoder operated at the average bitrate of 8 kb/s. There were 10 expert listeners in the test. The test results are shown in Fig. 3a, plotted along with the 95% confidence intervals (normal distribution). It can be seen that the sRNN decoder provides performance comparable to SILK at half of its bitrate.

In a second experiment, we used an average bitrate of 6.4 kb/s for the vocoder and sRNN. The results of the test are shown in Fig. 3b. There were 11 expert listeners taking part in this test. It can be seen that the 6.4 kb/s sRNN provides performance better than AMR-WB at 23.05 kb/s, but it remains in a similar quality region. However, it can also be seen that sRNN at 6.4 kb/s is significantly worse than SILK. We note that the two tests were carried out in separate sessions. The listener populations were only partially overlapped between the two tests.

In addition, we evaluated the performance of the conditions from the listening test using the objective quality assessment tool POLQA [29]. The results are shown in Table 2. Although this tool is not suitable for evaluation of signals synthesized by non-deterministic generative models, it is interesting to note a relatively high score for the vocoded conditions, and relatively low scores for sRNN. These results deviate significantly from the results of our listening tests. However, the ordering of the classic codecs, AMR-WB and SILK is the same as in our listening tests.
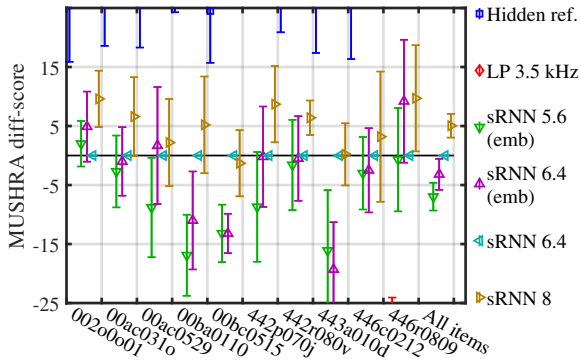
Fig. 4. The differential scores in the rate-quality test of the embedded sRNN decoders with respect to the sRNN trained on the 6.4 kb/s conditioning (95% confidence intervals, Student-t distribution.)
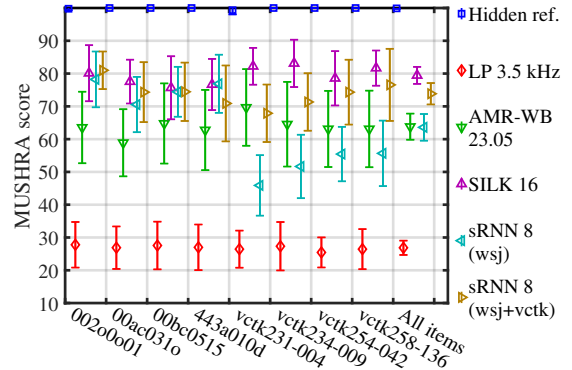

Fig. 5. The results of listening test including sRNN decoders operating at 8 kb/s trained exclusively on WSJ0 and combination of WSJ0 and VCTK; there were 9 expert listeners.

4.2. Quality-bitrate trade-off

In a third experiment, we evaluated the perceptual quality-bitrate trade-off achieved with the sRNN decoder. The conditions in this test included sRNN decoders evaluated in Section 4.1. In addition, we took an 8 kb/s decoder and used it to decode bitstreams generated at lower rates (5.6 kb/s and 6.4 kb/s) – without retraining. We used properties of the conditioning that allowed to embed a lower rate conditioning into a higher rate conditioning and provide this to a decoder trained on higher-rate conditioning (see Section 2).

The results of the experiment are shown in Fig. 4. There were 10 expert listeners in the test. It can be seen that, for the provided data points, sRNN 8 decoder provides a graceful quality degradation with decreasing bit rate (sRNN 6.4 (emb), sRNN 5.6 (emb)). Furthermore, the performance of the decoder using the embedded conditioning (sRNN 6.4 (emb)) is similar to the performance of a decoder trained for the specific bitrate (sRNN 6.4). While we cannot conclude that the result generalizes to a wide range of bitrates, we can see that the proposed sRNN decoder learned to generalize the conditioning, which can be used to provide meaningful rate-distortion trade-offs. We note that in the course of the training the 8 kb/s decoder was never exposed to lower quality conditioning.

4.3. Perceptual evaluation outside the WSJ0 set

We found no evidence of model overfitting in our experiments within the WSJ0 dataset and all evaluation was done on recordings not included in training. However, while evaluating the coding scheme on speech signals outside WSJ0 we found a performance drop in order of 10-15 MUSHRA points on average. In order to evaluate this effect in a controlled experiment, we used signals from another publicly available speech database.

We constructed a new dataset based on roughly equal contributions from the WSJ0 set and VCTK set (clean speech from [18]), and also TSP set [30](1% of the full data set). The final set composition resulted in doubling the total number of speakers. The total recording time of the combined set was still similar to WSJ0 in the previous

experiments. The training set comprised 45 hours of speech and we used 10 hours of speech for the validation set and 10 hours for the test set. There was no speaker overlap between the three sets and the material was approximately equally balanced between female and male speakers. We retrained the sRNN decoder on this new dataset.

For our final experiment, we randomly selected items from WSJ0 and VCTK test sets (4 female speakers, 4 male speakers), and used two versions of identically configured sRNN decoders (8 kb/s). One was trained solely on WSJ0, the other one was trained on the combination of the datasets as described above. We also used the two classic speech codecs as the quality anchors configured identically as in the previous experiment. The listening test was conducted with 9 expert listeners and its mean results are shown in Fig. 5 together with 95% confidence intervals (normal distribution).

It can be seen in the results that the sRNN trained on WSJ0 has significantly worse performance on the items coming from the VCTK set. But the performance of the retrained sRNN is similar to the performance reported in Section 4.1. We note that the performance on the WSJ0 items has been maintained while the performance on the signals from the updated data set has improved significantly. The main difference between the WSJ0 items and the VCTK items is that the latter comprise dry speech captured in a semi-anechoic chamber.

We have also informally evaluated the scheme on material in Mandarin, French, Swedish, and German and we did not find a performance drop due to operation on speech in these languages. Note that all the training material was in English.

5. CONCLUSIONS

We demonstrated that it is possible to achieve high quality speech coding that is on par with the state-of-the-art speech codecs, but at much lower bitrates. It is possible in a coding scheme, where quantized vocoder parameters are used to condition the SampleRNN. Furthermore, we demonstrated that this architecture has a potential to provide a rate-quality trade-off, which we established by conducting a series of formal listening tests.

We found the performance of such schemes is highly dependent on the composition of the datasets and that the robustness issues can be addressed by using more diverse training material.

Table 2. Average POLQA scores for conditions and items from the listening tests (wide-band setting)

	Vocoder		AMR-WB	SILK	sRNN	
Rate [kb/s]	6.4	8.0	23.05	16.0	6.4	8.0
MOS-LQO	3.43	3.67	4.39	4.41	3.27	3.48

6. REFERENCES

[1] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu, "WaveNet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016.

[2] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio, "SampleRNN: An unconditional end-to-end neural audio generation model," *arXiv preprint arXiv:1612.07837*, 2016.

[3] Sercan O Arik, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, et al., "Deep voice: Real-time neural text-to-speech," *arXiv preprint arXiv:1702.07825*, 2017.

[4] Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio, "Char2wav: End-to-end speech synthesis," *ICLR workshop*, 2017.

[5] Antonio Bonafonte, Santiago Pascual, and Georgina Dorca, "Spanish statistical parametric speech synthesis using a neural vocoder," in *Proc. Interspeech 2018*, 2018, pp. 1998–2001.

[6] Yang Ai, Hong-Chuan Wu, and Zhen-Hua Ling, "SampleRNN-based neural vocoder for statistical parametric speech synthesis," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5659–5663.

[7] Cong Zhou, Michael Horgan, Vivek Kumar, Cristina Vasco, and Dan Darcy, "Voice Conversion with Conditional SampleRNN," in *Proc. Interspeech 2018*, 2018, pp. 1973–1977.

[8] Per Hedelin, "A sinusoidal LPC vocoder," in *2000 IEEE Workshop on Speech Coding. Proceedings. Meeting the Challenges of the New Millennium (Cat. No.00EX421)*, Sept 2000, pp. 2–4.

[9] B. Bessette, R. Salami, R. Lefebvre, M. Jelinek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, and K. Jarvinen, "The adaptive multirate wideband speech codec (AMR-WB)," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 8, pp. 620–636, Nov 2002.

[10] Soeren Jensen, Koen Vos, and Karsten Soerensen, "SILK speech codec," *Working Draft, IETF Secretariat, Fremont, CA, USA, Tech. Rep. draft-vos-silk-02. txt*, 2010.

[11] Anssi Rämö and Henri Toukomaa, "Voice quality evaluation of recent open source codecs," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[12] W. B. Kleijn, F. S. C. 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)*, April 2018, pp. 676–680.

[13] "Method for the subjective assessment of intermediate quality levels of coding systems," Recommendation ITU-R BS.1534-3, Oct. 2015.

[14] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma, "Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications," *arXiv preprint arXiv:1701.05517*, 2017.

[15] Andrew Ng, "Nuts and bolts of deep learning," *NIPS 2016, Bay Area Deep Learning School*, 2016.

[16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning*, MIT Press, 2016.

[17] John S. Garofolo and et al., "CSR-I (WSJ0) complete LDC93S6A," Philadelphia: Linguistic Data Consortium, 1993.

[18] Cassia Valentini-Botinhao et al., "Noisy speech database for training speech enhancement algorithms and TTS models," University of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR), 2017.

[19] Allen Gersho and Robert M Gray, *Vector quantization and signal compression*, vol. 159, Springer Science & Business Media, 2012.

[20] Kuldip K Paliwal and Bishnu S Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *J. Acoust. Soc. Am.*, vol. 87, no. 1, pp. S39, 1990.

[21] Turaj Zakizadeh Shabestary, Per Hedelin, and Fredrik Nordén, "Entropy constrained quantization of LSP parameters," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[22] Turaj Zakizadeh Shabestary and Per Hedelin, "Vector quantization by companding a union of z-lattices," *IEEE transactions on information theory*, vol. 51, no. 2, pp. 738–746, 2005.

[23] Jan Lindén, Jan Skoglund, and Thomas Eriksson, "Improving predictive vector quantizers in speech coding applications," in *Proc. Conf. NORSIG, Helsinki, Finland*. NORSIG, 1996.

[24] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[25] Aaron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George van den Driessche, Edward Lockhart, Luis C Cobo, Florian Stimberg, et al., "Parallel WaveNet: Fast high-fidelity speech synthesis," *arXiv preprint arXiv:1711.10433*, 2017.

[26] Ronald J Williams and Jing Peng, "An efficient gradient-based algorithm for on-line training of recurrent network trajectories," *Neural computation*, vol. 2, no. 4, pp. 490–501, 1990.

[27] Ilya Sutskever, *Training recurrent neural networks*, Ph.D. thesis, University of Toronto, 2013.

[28] Diederik P Kingma and Jimmy Ba, "ADAM: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[29] "Perceptual objective listening quality assessment," Recommendation ITU-T P.863, Jan. 2011.

[30] Peter Kabal, "TSP Speech Database," Telecommunications & Signal Processing Lab, Dept. of Elec. & Computer Engineering, McGill University, Sept. 2002.