WAVENET BASED LOW RATE SPEECH CODING

W. Bastiaan Kleijn,[1,3] *Felicia S. C. Lim,*[1] *Alejandro Luebs,*[1] *Jan Skoglund,*[1]
Florian Stimberg,[2] *Quan Wang,*[1] *Thomas C. Walters*[2]

[1]Google Inc., San Francisco, CA; [2] DeepMind, London, UK; [3]Victoria University of Wellington, NZ

ABSTRACT

Traditional parametric coding of speech facilitates low rate but provides poor reconstruction quality because of the inadequacy of the model used. We describe how a WaveNet generative speech model can be used to generate high quality speech from the bit stream of a standard parametric coder operating at 2.4 kb/s. We compare this parametric coder with a waveform coder based on the same generative model and show that approximating the signal waveform incurs a large rate penalty. Our experiments confirm the high performance of the WaveNet based coder and show that the speech produced by the system is able to additionally perform implicit bandwidth extension and does not significantly impair recognition of the original speaker for the human listener, even when that speaker has not been used during the training of the generative model.

Index Terms— Speech coding, parametric coding, WaveNet, generative model

1. INTRODUCTION

Speech coding found its first major application in secure communications, e.g., [1], and later enabled low-cost mobile and internet communications, e.g., [2, 3, 4]. With the continuously decreasing cost of bandwidth in most applications, the trade-off between rate and quality has gravitated to higher rates, typically over 16 kb/s, to ensure good quality. This state-of-the-art may change with the advent of a new generation of coders that provides a significant leap in performance. In this paper we discuss an approach that can provide good quality at rates around 2-3 kb/s with significant potential for further improvement in the rate-quality trade-off.

The redundancy in rate of existing speech coders can be determined from estimates of the information rate in speech. A recent rate estimate [5] based on comparing signals with the same message is consistent with lexical information rates computed from phoneme statistics [6]. They suggest that the true information rate is less than 100 b/s. Attributes of speech that identify the speaker and speaking style do not vary rapidly over time and hence do not change this rough estimate significantly. The common coding algorithms used in current communication systems require a rate that is roughly two orders of magnitude higher than the rate of the information conveyed.

Essentially, all speech coding methods are based on an explicit model of the signal, which usually is time varying. In *parametric coding* the signal is generated at the decoder based on the model parameters only. The quality of the signal reproduced by parametric coders is limited by the efficacy of the model. However, even poor signal models can be usefully exploited for high-quality coding. *Waveform coding* exploits that conditioning information (the model and its parameters) reduces the minimum rate required to achieve a particular mean error for the signal waveform. The penalty paid is that the reproduced signal is an approximation of the original waveform. This requires the transmission of information that, at least in principle, is not needed for high fidelity reconstruction, explaining in part the high rate of current speech coding schemes.

Most models used in speech coding have a statistical basis. A speech signal can be described as a discrete-time stochastic process $\{s_i\}$ with a non-zero (differential) entropy rate. A discrete-time stochastic process can be characterized by a sequence of conditional probability density functions (PDFs) $f(s_i|s_{i-1}, s_{i-2}, \cdots)$. If the memory is p samples, then the process is an order-p Markov processs. Application of the chain rule relates the PDF of a sample sequence to that of the conditional PDFs.

Various generative models have been used to describe the speech process. The models generally assume speech to be a Markov process and provide a conditional distribution for the next signal sample given a set of past samples. Ubiquitous are linear autoregressive (AR) models [7], and hidden Markov models (HMM) [8]. Refined generative models such as ARMA, e.g., [9], and kernel density estimation (KDE) HMM, e.g., [10], can also be used. However, linear AR modeling remains the most commonly used generative model in speech coding. This may change with the recent introduction of the deep neural network (DNN) based WaveNet [11] generative models.

Recursive sampling of the conditional PDF of a speech model can be used to produce a speech signal. The sampling of the PDF corresponds to the generation of new information. To retain the perceptible attributes of an original speech signal in the coding application, the PDF must include conditioning variables (side information). These typically specify the short-term power spectral density, pitch and periodicity level. Good signal quality can be guaranteed even for imperfect generative models by approximating the original signal waveform. However, no new information is then generated during reconstruction and this information must be transmitted instead. Efficiency can be increased by encoding of blocks of samples simultaneously, for example using analysis-by-synthesis [12] in the context of generative models. The analysis-by-synthesis paradigm applied to AR models, introduced in [13], is universally used in mobile phone standards.

The contribution of this paper is the usage of WaveNet as a generic generative model for speech coding and an analysis thereof. We describe two WaveNet based coding architectures: *i*) a parametric coder that encodes only the conditioning variables, and *ii*) a waveform coder that encodes the conditioning variables and also the observed waveform exploiting the conditional distribution. The parametric coder model differs from [14] in that it is not speaker dependent, and that it can also be decoded with a conventional low complexity decoder. Our coding architectures replace the text sequence used as conditioning information in the original text-to-speech (TTS) application of WaveNet by a sequence of quantized parameters of a parametric speech coder. This change is nontrivial as in a coding scenario the generative model cannot be trained on the speakers that

the coder encounters during operation. It will be shown in section 3 that WaveNet can be generalized to the multi-speaker case.

In the remainder of this paper, we first describe our approach in more detail in section 2, then discuss our experimental results in section 3 and finally provide conclusions in section 4.

2. ALGORITHM

In this section we discuss the parametric coding architecture, analyze the rate and describe the waveform coding architecture.

2.1. Parametric WaveNet Coder

A parametric coder transmits only the conditioning variables of the generative model that generates the signal at the decoder. It is possible to train a neural architecture at the encoder to optimize the conditioning variables. In this paper, we opted for a conventional parametric encoder instead. The latter approach has a significantly lower complexity at the encoder and facilitates the use of a low complexity decoder if computational resources fall short. We first motivate our conventional encoder choice and then describe the WaveNet decoder.

Traditional parametric coders almost always encode a similar set of parameters: spectral envelope, pitch, and voicing level. The parameter set differs little for approaches based on a temporal perspective with glottal pulse trains [15, 16], and those based on a frequency-domain perspective with sinusoids [17, 18]. Any of these parameter sets can be used as a set of conditioning variables for WaveNet.

To illustrate that our architecture does not require special features for the encoder, we selected Codec 2 [19], an open source speech coder. It belongs to the sinusoidal coder family and can run at various update rates. For example, at 2.4 kb/s, each 20 ms block encodes the short-time spectral envelope using line spectral frequencies [20] with 36 bits, the pitch with 7 bits, the signal power with 5 bits, and the voicing level with 2 bits. The voicing level is determined as in the multiband excitation vocoder [21].

We note that parametric coders (including Codec 2) almost universally operate on narrow-band speech with a sampling rate of 8 kHz but for high quality output speech, a wide-band signal ($\geq$ 16 kHz) is preferred. Historically, wide-band extension is optionally applied after the decoder [22]. In our approach, we train our decoder with 8 kHz conditioning variables and 16 kHz speech signals such that it implicitly performs bandwidth extension.

For the parametric decoder, we use the WaveNet generative model [11]. Given the past output signal and the conditioning variables, WaveNet provides a discrete probability distribution of the next signal sample using the 8-bit ITU-T G.711 μ-law format. It then samples this distribution to select the output sample value.

The WaveNet architecture is a multi-layer structure using dilated convolution with gated cells. The conditional variables are supplied to all layers of the network. For the coder, we retained the standard WaveNet configuration of [11] but replaced the conditioning variables with the decoded Codec 2 bit stream. The Codec 2 decoder provides its parameters to its sinusoidal renderer at 100 Hz, which we used unchanged as the conditioning variables for the WaveNet decoder. As WaveNet requires conditioning for each output sample, we hold the conditional variables constant for 10 ms intervals.

During training, WaveNet learns the parameters of a softmax function that represents the conditional discrete probability distribution. The training is subject to the same conditioning variables that are used during run-time. In contrast to WaveNet training for TTS applications, we used a training database containing a large number of different talkers providing a wide variety of voice characteristics, all without conditioning on a label that identifies the talker.

2.2. Rate Analysis

The rate benefit of *generating* the waveform over approximating the original waveform can be estimated from the information rate generated at the decoder. Basic information theory allows us to estimate the relevant rates. Let us consider a generated signal sequence $\{S_i\}_{i\in\mathcal{A}}$, where $\mathcal{A}$ is an index sequence, and a conditioning sequence $\{\Theta_i\}_{i\in\mathcal{A}}$, with the two sequences being presented at the same sampling rate. Let both sequences be discrete valued and of finite length $|\mathcal{A}|$. Their entropy rates then satisfy

$$\frac{1}{|\mathcal{A}|}H(\{S_i\},\{\Theta_i\}) = \frac{1}{|\mathcal{A}|}H(\{S_i\}|\{\Theta_i\}) + \frac{1}{|\mathcal{A}|}H(\{\Theta_i\}), \quad (1)$$

where we omitted sequence subscripts to simplify notation. Hence, the overall information rate contained in the decoded signal is the sum of the information rate associated with the generative process and the rate of the encoded conditioning parameters. The rate required for the conditioning variables is upper bounded by the rate of the parametric coder.

The information rate $\frac{1}{|\mathcal{A}|}H(\{S_i\}|\{\Theta_i\})$ associated with the generative process is simple to evaluate for WaveNet. We consider $|\mathcal{A}| \to \infty$ and make the additional assumption that speech is a short-time stationary and ergodic process and that, therefore, $\{S_i\}$ and $\{\Theta_i\}$ are stationary.

At its output, the WaveNet decoder produces a probability distribution with a set $\mathcal{N}$ of 256 discrete values for a μ-law encoding of the next sample. We denote this distribution as $q_n^{(i)}$, $i \in \mathbb{Z}$, $n \in \mathcal{N}$. The distribution $q^{(i)}$ is sampled to produce the scalar output signal value. The mean information in bits generated by this sampling operation is the conditional entropy of the distribution:

$$H(S_i|s_{i-1}, s_{i-2}, \cdots; \theta_i) = -\sum_{n\in\mathcal{N}} q_n^{(i)} \log_2 q_n^{(i)}, \quad (2)$$

where θ_i is the current vector of conditioning variables, and where capital letters indicate random variables and lower-case letters realizations. Note that $H(S_i|s_{i-1}, s_{i-2}, \cdots; \theta_i)$ is simple to compute and that the evaluation is exact for the reconstruction process.

We can now find the generated signal rate using an approximation to the chain rule:

$$\lim_{|\mathcal{A}|\to\infty} \frac{1}{|\mathcal{A}|}H(\{S_i\}|\{\Theta_i\}) = H(S_i|S_{i-1}, S_{i-2}, \cdots; \Theta_i) \quad (3)$$

$$\approx \frac{1}{|\mathcal{A}_0|}\sum_{i\in\mathcal{A}_0} H(S_i|s_{i-1}, s_{i-2}, \cdots; \theta_i), \quad (4)$$

where $\mathcal{A}_0$ is an observed finite sequence. We will evaluate the result in section 3.

2.3. WaveNet Waveform Coder

As will be seen in section 3, relation (4) implies that a coder that reproduces the signal waveform is inefficient compared to a coder relying on a generative model. However, a generative coder cannot guarantee its output quality; situations may occur where the modelled generative distribution is not a good description of the signal.

Importantly, scenarios where the parametric WaveNet coder does not perform well can be detected within the WaveNet framework. Running a waveform WaveNet at the encoder allows the evaluation of the log likelihood of the input signal based on the WaveNet

generative model and compare that to the expectation. Thus, we can select between parametric and waveform coding (low and high rate). When generative performance is good, the decoder receives no waveform information, and it reverts to a generative mode. Resynchronization can be achieved with conventional techniques [23]. We will report on this mode selection paradigm elsewhere and only discuss the basic WaveNet waveform coder here.

Waveform coding is commonly based on prediction, particularly at rates below 30 kb/s. While fixed-rate predictive coding is most common, variable-rate versions also exist. In predictive coders, the conditional probability distribution is generally considered fixed relative to its mean predicted value. The same approach can be used for WaveNet coding, but it is not natural. In WaveNet waveform coding, the prediction step can be omitted as a conditional discrete distribution of the next sample is available without actually computing a predicted sample value.

We now describe our variable-rate WaveNet waveform coding structure. We first discuss the quantization step and the corresponding decoding step. Let $Q : \mathbb{R} \to \mathcal{N}$ be the mapping from the signal to its quantization index and $Z : \mathcal{N} \to \mathbb{R}$ be the corresponding decoding operation. Then, a signal x_i is encoded at the encoder as $n_i = Q(x_i)$ and the quantized signal is obtained at the decoder as $\hat{x}_i = Z(n_i)$. The resolution of Q determines both the rate and the quality for the waveform coder. In conventional WaveNet, Q is a μ-law quantizer. Our proposed approach described below will therefore result in the exact same waveform as basic μ-law coding.

In the WaveNet waveform coder, the sequence of quantization indices $\{n_i\}_{i\in\mathbb{Z}}$ is subject to entropy coding and subsequently transmitted over the channel along with the conditioning variables. Identically trained WaveNet models are required at both the encoder and decoder to provide the conditional probability distribution $q^{(i)}$ to the entropy encoder and decoder. Importantly, it is $\hat{x}_i$ that is used at the decoder as input for generating subsequent samples in the WaveNet model, thus creating a *closed loop* coder. (In contrast, the parametric WaveNet coder takes the previous generated sample as input.) Although $q^{(i)}$ is an approximate distribution for the original signal $\{x_i\}_{i\in\mathcal{A}}$, it can be used to reduce the rate significantly with entropy coding. Using techniques such as arithmetic coding [24, 25] can provide near-optimal coding for an entire sequence of indices. If the predictive distributions $q^{(i)}$ are correct, then the rate of the waveform coder will be arbitrarily close to that of (4). Any mismatch in the conditional distribution results in an expected rate increase that is specified by the Kulback-Leibler divergence.

The described WaveNet waveform coding scheme is close to optimal for the squared error measure on samples with a μ-law warped amplitude. The cubic cell shape of scalar quantization imposes a penalty that is (asymptotically with increasing rate) maximally 1.5 dB in mean squared error distortion or, equivalently, 0.25 bits per sample [26]. As conditional distributions for higher dimensionalities and the usage of analysis-by-synthesis suffer from high complexity, the removal of this penalty is non-trivial.

The WaveNet waveform coder can be characterized by two rates. On the one hand we can compute the average entropy rate of the estimated conditional distribution over the observed sequence:

$$\bar{H} = -\frac{1}{|\mathcal{A}_0|} \sum_{i\in\mathcal{A}_0} \sum_{n\in\mathcal{N}} q_n^{(i)} \log_2 q_n^{(i)}, \tag{5}$$

which provides an estimate of the true average entropy rate. On the other hand, we can compute a lower bound on the real-world rate produced by the entropy coder over the observed sequence:

$$R = -\frac{1}{|\mathcal{A}_0|} \sum_{i\in\mathcal{A}_0} \log_2 q_{n_i}^{(i)}. \tag{6}$$

which forms an upper bound on the average entropy rate. If R and $\bar{H}$ are close, we can be confident that the true average entropy rate is similar. The measured rates will be discussed in section 3.

Finally, we discuss the rate required for the conditioning variables of a WaveNet waveform coder. It was shown in [27] that under reasonable conditions *the optimal rate for the conditioning variables does not depend on the mean signal distortion*. This also applies to the WaveNet based waveform coder and implies that only the resolution of the quantizer Q has to be adjusted to vary the rate.

We did not include perceptual weighting in the current implementation of the waveform WaveNet coder. However, pre- and post-filtering structures can be introduced to enhance coder performance by exploiting perception.

3. EXPERIMENTAL RESULTS

In this section, we evaluate the information rates, signal quality and speaker identifiability produced by the WaveNet coding schemes. Listening examples are available online.[1]

3.1. Experimental Setup

The WaveNet system as described in [11] was used to develop our proposed WaveNet coders. At the encoder, we employed Codec 2 at 8 kHz and 2.4 kb/s. As previously discussed, its decoded bit stream was used as WaveNet conditioning variables, held constant over 10 ms intervals. The decoder then generated output samples at 16 kHz.

The training and test sets were derived from the Wall Street Journal speech corpus [28] with no overlapping speakers. The training set contained 32580 utterances by 123 speakers and the testing set contained 2907 utterances by 8 speakers.

Quality was evaluated using POLQA and listening tests against a number of unmodified reference coders: Codec 2 (2.4 kb/s), MELP (2.4 kb/s) [29], Speex wideband (2.4 kb/s) [30], ITU-T G.711 μ-law at 16 kHz (128 kb/s), and ITU-T G.722.2 AMR-WB (23.05 kb/s).

Speaker identification performance of the parametric WaveNet coder was evaluated with listening tests and a neural network based model [31] trained on our dataset. As reference, we trained a second parametric WaveNet coder with overlapping speakers in the training and test sets.

3.2. Speech Information Rates

We used (4) to estimate the rate of waveform generation on the test dataset (removing silence segments). The result was a mean rate of 2.65 bits per *sample*, or 42 kb/s at 16 kHz. Informal testing indicates that this rate is largely independent of the rate of the conditioning parameters. We also estimated the waveform-coding rates associated with (5) and (6) and obtained 2.61 and 2.62 bits per sample respectively, or about 42 kb/s at 16 kHz. Since they are very similar, we can be confident that the true average entropy rate is similar. Fig. 1 shows a speech segment of 0.2 seconds and the corresponding *instantaneous* information rates for the waveform coder. The early part of the signal is a fricative, which is relatively unstructured, thus a higher rate is required. The latter part is a voiced segment, where the rate required is low. The rate additionally varies with the pitch

[1] https://goo.gl/C14FFx

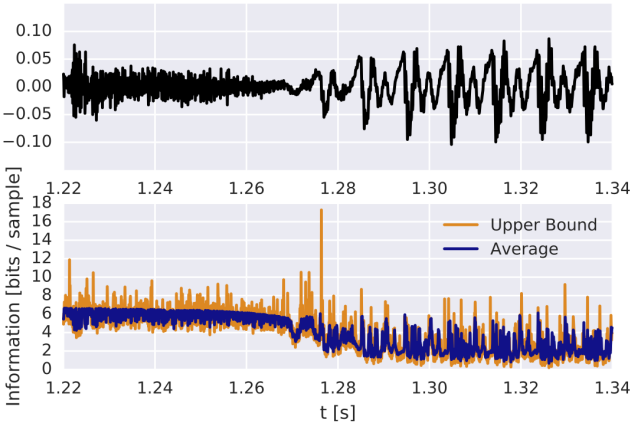

Fig. 1: Top: signal waveform. Bottom: instantaneous information computed as the terms in the sample average sums in (5) (blue) and (6) (orange).

cycle. The highest rate is not associated with the pitch pulse, implying that WaveNet predicts the pitch accurately, but the waveform is noisier across certain segments of the pitch cycle.

3.3. Quality Experiments

The objective quality of the reference and two WaveNet coders was evaluated using POLQA [32] and the results are shown in Table 1. We noted that POLQA did not reflect informal listening impressions, where the bandwidth extension and absence of the distortions typical of a vocoder-based parametric coder was clearly heard. This discrepancy was not unexpected as the parametric WaveNet coder changes the signal waveform and the timing of the phones.

A subjective MUSHRA-type listening test [33] was performed where 21 participants (1/3 of the total were experts) evaluated eight utterances. The μ-law coder was omitted from the test as it is identical to the WaveNet waveform coder. The results are given in Fig. 2 where it can be seen that two distinctive groups emerged: a low-quality group consisting of Speex, Codec 2 and MELP, and a high-quality group consisting of AMR-WB, the WaveNet waveform coder, and the parametric WaveNet coder. Thus, the parametric WaveNet coder has a subjective quality similar to that of existing waveform coders with the benefit of significantly lower rates.

3.4. Speaker Identification Experiments

An objective speaker identification test was performed using a neural network based speaker identification model [31]. Two single-layer models were trained with μ-law coded and parametric WaveNet coded speech respectively. There were no overlapping speakers between the training and test data. The training set contained 123 speakers and 3690 utterances and the same test set as before was used, but now split into enrollment and verification sets with overlapping speakers but non-overlapping utterances. The verification equal error rate (EER) results were 8.4% for the μ-law coded speech and 15.8% for WaveNet coded speech. It is known that EER goes up after coding [34] and in this case, it is expected that the spectral resolution of the low rate coder has restricted speaker identifiability.

A listening test was also carried out. For this, a second parametric WaveNet coder was trained with overlapping speakers between the training and test data. We term this model as W_w and the first model trained without overlapping speakers as W_{wo}. A triangle test was performed where 15 listeners listened to 16 trials, each with three different utterances by the same speaker. Two utterances were taken from one of $\{W_w, W_{wo}\}$ and one utterance from the other model. The total utterances from each model were equal over all trials. The listeners had to indicate the utterance spoken by the different speaker. On average, they correctly identified the different speaker in 41% of the trials. If the speakers are indistinguishable, the expected value is 33%. This discrepancy is likely to decrease with an increasing number of speakers in the training set.

These experiments indicate that the current coder would be more suitable for human-facing applications, e.g., conference calls.

Table 1: POLQA mean opinion scores (MOS-LQO) for different coders operating at different rates (kb/s). WW: WaveNet waveform coder, WP: parametric WaveNet coder.

	Codec 2	*MELP*	*Speex*	*AMR-WB*	*WW*	*WP*
Rate	2.4	2.4	2.4	23	42	2.4
MOS	2.7	2.9	2.2	4.6	4.7	2.9

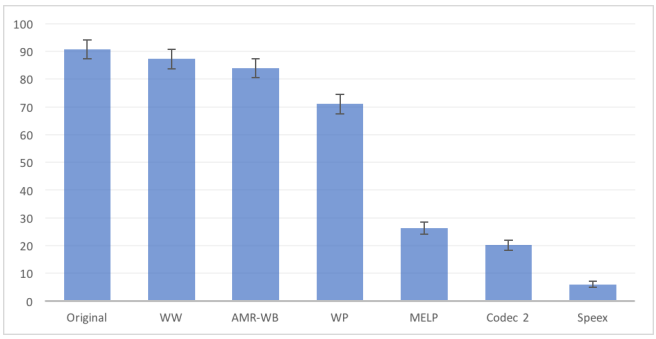

Fig. 2: Subjective quality (MUSHRA scores). WW: WaveNet waveform coder, WP: parametric WaveNet coder. The error bars show 95% confidence intervals.

4. CONCLUSIONS

Our results show that the high fidelity of the conditional probability distribution of the speech waveform of WaveNet can be leveraged to create state-of-the-art speech and, likely, audio coding systems. We demonstrated that the quality of our 2.4 kb/s parametric speech coder is similar to that of waveform coders with much higher rates. We also showed how to build a waveform WaveNet coder and briefly discussed a method for switching from the parametric coder to the waveform coder based on a likelihood based quality measure for the parametric coder.

The computational cost of both training and running WaveNet is high compared to conventional coders. The exception is the parametric WaveNet encoder, which is a conventional, low complexity parametric encoder.

It is expected that performance can be improved further. For example, the conditioning parameter set and their interpolation over time can be further refined and pre-/postfiltering can be introduced to the waveform coder to improve perceived performance. To increase computational efficiency, it may be beneficial to study if the long-lag memory components of the conditional probability distribution unnecessarily duplicates the information that is also present in the bit stream.

5. REFERENCES

[1] T. Tremain, "Linear predictive coding systems," in *ICASSP '76. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, Apr 1976, pp. 474–478.

[2] P. Kroon, E. Deprettere, and R. Sluyter, "Regular-pulse excitation–a novel approach to effective and efficient multipulse coding of speech," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 34, no. 5, pp. 1054–1063, Oct 1986.

[3] R. Salami, C. Laflamme, J. P. Adoul, and D. Massaloux, "A toll quality 8 kb/s speech codec for the personal communications system (PCS)," *IEEE Transactions on Vehicular Technology*, vol. 43, no. 3, pp. 808–816, Aug 1994.

[4] W. B. Kleijn, R. P. Ramachandran, and P. Kroon, "Interpolation of the pitch-predictor parameters in analysis-by-synthesis speech coders," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 1, pp. 42–54, Jan 1994.

[5] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "On the information rate of speech communication," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 5625–5629.

[6] P. B. Denes, "On the statistics of spoken English," *J. Acoust. Soc. Am.*, vol. 35, no. 6, pp. 892–904, 1963.

[7] B. S. Atal and M. R. Schroeder, "Adaptive predictive coding of speech signals," *The Bell System Technical Journal*, vol. 49, no. 8, pp. 1973–1986, Oct 1970.

[8] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, vol. 3, 2000, pp. 1315–1318.

[9] Y. Grenier, "Time-dependent ARMA modeling of nonstationary signals," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 31, no. 4, pp. 899–911, Aug 1983.

[10] M. Piccardi and Ó. Pérez, "Hidden Markov models with kernel density estimation of emission probabilities and their use in activity recognition," in *Comp. Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[11] A. van den 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 e-prints*, Sep. 2016.

[12] C. Bell, H. Fujisaki, J. Heinz, K. Stevens, and A. House, "Reduction of speech spectra by analysis-by-synthesis techniques," *J. Acoust. Soc. of America*, vol. 33, no. 12, pp. 1725–1736, 1961.

[13] S. Singhal and B. Atal, "Improving performance of multi-pulse LPC coders at low bit rates," in *ICASSP '84. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 9, Mar 1984, pp. 9–12.

[14] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder," in *Proceedings Interspeech*, 2017.

[15] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. America*, vol. 50, no. 2B, pp. 637–655, 1971.

[16] J. D. Markel and A. J. Gray, *Linear prediction of speech*. Springer Verlag, 1976, vol. 12.

[17] P. Hedelin, "A tone oriented voice excited vocoder," in *ICASSP '81. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 6, Apr 1981, pp. 205–208.

[18] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 34, pp. 744–754, 1986.

[19] D. Rowe. (2011) Codec 2- open source speech coding at 2400 bits/s and below. [Online]. Available: http://www.tapr.org/pdf/DCC2011-Codec2-VK5DGR.pdf

[20] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals," *J. Acoust. Soc. Am.*, vol. 57, p. 535(A), 1975.

[21] 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, Aug 1988.

[22] B. Iser, G. Schmidt, and W. Minker, *Bandwidth Extension of Speech Signals*. Springer, 2008.

[23] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *Acoustics, Speech, Signal Process., ICASSP-93, IEEE Int. Conf. on*, vol. 2, 1993, pp. 554–557.

[24] R. Pasco, "Source coding algorithms for fast data compression," Ph.D. dissertation, Stanford University, 1976.

[25] J. J. Rissanen and G. Langdon, "Arithmetic coding," *IBM J. Res. Devel.*, vol. 23, no. 2, pp. 149–162, 1979.

[26] T. Lookabough and R. Gray, "High-resolution theory and the vector quantizer advantage," *IEEE Trans. Information Theory*, vol. IT-35, no. 5, pp. 1020–1033, 1989.

[27] W. B. Kleijn and A. Ozerov, "Rate distribution between model and signal," in *Proc. IEEE Workshop on Applic. Signal Process. Audio and Acoust. (WASPAA)*, Nov. 2007, pp. 243–246.

[28] Eugene Charniak et al., "BLLIP 1987-89 WSJ corpus release 1 ldc2000t43 web download," Linguistic Data Consortium, Philadelphia, 2000.

[29] A. McCree, K. Truong, E. George, T. Barnwell, and V. Vishu, "A 2.4 kbit/s MELP coder candidate for the new U.S. federal standard," in *Acoustics, Speech, Signal Process., 1988. ICASSP-88., 1988 Int. Conf. on*, vol. 1, 1996, pp. 200 – 203 vol. 1.

[30] J. M. Valin. (2016) Speex. [Online]. Available: https://www.speex.org

[31] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *Submitted to ICASSP 2018*, 2018.

[32] ITU, "Perceptual objective listening quality assessment," *Rec. ITU-T P.863*, 2011.

[33] ——, "Method for the subjective assessment of intermediate sound quality (MUSHRA)," *Rec. ITU-R.BS.1534-1*, 2003.

[34] R. B. Dunn, T. F. Quatieri, D. A. Reynolds, and J. P. Campbell, "Speaker recognition from coded speech and the effects of score normalization," in *Conference Record of Thirty-Fifth Asilomar Conference on Signals, Systems and Computers (Cat.No.01CH37256)*, vol. 2, Nov 2001, pp. 1562–1567 vol.2.