ON THE USE OF WAVENET AS A STATISTICAL VOCODER

Nagaraj Adiga, Vassilis Tsiaras, Yannis Stylianou

Department of Computer Science, University of Crete, Greece
nagaraj@csd.uoc.gr, tsiaras@csd.uoc.gr, yannis@csd.uoc.gr

ABSTRACT

In this paper, we explore the possibility of using the WaveNet architecture as a statistical vocoder. In that case, the generation of speech waveforms is locally conditioned only by acoustic features. Focusing on the single speaker case at the moment, we investigate the impact of the local conditions as well as that of the amount of data available for training. Furthermore, variations of the WaveNet architecture are considered and discussed in the context of our work. We compare our work against a very recent work which also used WaveNet architecture as a speech vocoder using the same speech data. More specifically, we used two female and two male speakers from the CMU-ARCTIC database to contrast the use of cepstrum coefficients and filter-bank features as local conditioners with the goal to improve the overall quality for both male and female speakers. In the paper we also discuss the impact of the size of the training data. Objective metrics for quality and intelligibility of the generated by the WaveNet speech as well as subjective tests support our suggestions.

Index Terms— WaveNet, filter-bank features, vocoder, autoregressive model, correlated conditioning.

1. INTRODUCTION

Speech analysis and reconstruction usually involve the development of a mathematical representation of speech production mechanism. Parameters of such representation are estimated by minimizing an error criterion (usually a mean squared error) between the original speech and the mathematical model. Most of the mathematical representations suggested in the literature are based on the assumption of the source-filter theory for the production of speech. Sometimes the models are simplistic for modeling source and filter [1] or more elaborating [2, 3]. Speech models have found applications in speech coding as well as speech synthesis (text-to-speech). The capability to change the parameters of the model was found to be useful in modifying pitch, the time-scale of the signal [4–6], or even in modifying the identity of the speaker, like in voice conversion [7, 8]. Thus, these models have been found to be useful in various ways to synthesize speech: from diphone synthesizers to high-quality Unit Selection, or Hybrid, based speech synthesis systems as well for statistical based speech synthesis systems using HMM or DNN [9–12].

For all above applications of speech models, we usually consider between analysis and reconstruction steps, an intermediate step that of transformation of the parameters of the model. For speech coding that includes coding, transmission, decoding. In text-to-speech, that includes modification of model parameters. For example, in Unit-Selection based systems, for smoothing discontinuities during unit transitions. In Statistical approaches, for reconstructing/generating speech waveform from acoustic features predicted by linguistic information. Speech models are then classified as good or bad depending on their capability for being robust against these transformations, modifications, or predictions. For the two last cases, we expect speech models to generate speech that is as natural as possible, while in the first case (speech coding) we expect speech models to reconstruct speech as close as possible to the original speech.

From a machine learning point of view, speech models (although it depends on their degrees of freedom) provide an over fitting solution. Thus, slight modifications in the analysis parameters may generate artifacts during the reconstruction of the signal. For example, in voice conversion when a speech model uses the predicted acoustic features of the target speaker given those from a source speaker, the output quality has artifacts like buzziness. The same quality is also observed when the acoustic features are predicted from the linguistic information in the case of statistical text-to-speech systems. Recently, WaveNet [13] has been suggested for text-to-speech synthesis showing that a non-linear autoregressive system can mimic speech generation very well, while if it is appropriately locally conditioned with linguistic information, a high-quality text-to-speech synthesis system is obtained. If the local conditioning changes from linguistic information to acoustic information, then the WaveNet system is mainly a (statistical) vocoder [14]. The latter case has many applications for example in speech enhancement, speech modifications, voice Conversion [15–17] etc.

Inspired by these latest developments of WaveNet as a (statistical) vocoder, in this paper we are exploring the impact of conditioning as well as that of the amount of training data, in the quality of generated speech by WaveNet. To accelerate the speech training procedure, we consider a modified ver-

sion of the WaveNet as it was used in [14]. Using the same speech data as in [14] and with objective and subjective tests, we show that the choice of acoustic features as local conditioning affects the quality of the generated by the WaveNet speech.

The paper is organized as follows: In Section 2, details of WaveNet architecture explored in our work is explained. In Section 3, details our exploration of WaveNet architecture using mel-cesptrum coefficients and mel-filter-bank features as local conditioning is described. The experimental evaluations demonstrating the effectiveness of local conditioning and data variation experiments are presented in Section 4. Finally, the paper is concluded in Section 5.

2. WAVENET ARCHITECTURE

WaveNet is capable of synthesizing natural sounding speech especially if natural prosody is used [13]. The basic WaveNet is an autoregressive network, which generates a probability distribution of the next sample given some segment of previous samples. The next sample is produced by sampling from this distribution. An entire sequence of samples is produced by feeding previously generated samples back into the model. In order to make the training and generation tasks computationally tractable the discrete softmax is chosen as the probability distribution. Additionally, the dynamic range of speech samples are compressed via μ-law transformation and then quantized using 8-bits. The basic WaveNet produces babbling noise, which sounds like human speech if no conditioning is applied. In order to convey verbal and prosodic information the WaveNet is conditioned on linguistic and prosodic features [13, 14]. The local conditioning features are upsampled to the desired sampling frequency and fed into the basic WaveNet through a conditioning network. Let r be the receptive field of WaveNet, $x = \{x_1, x_2, \ldots, x_n\}$ be a sequence of quantized speech samples and $h = \{h_1, h_2, \ldots, h_n\}$ be the corresponding sequence of upsampled conditioning features. Assuming that $n > r$, the output of the conditioned WaveNet is described by the following conditional probability distribution.

$$P(x_n|x_{n-1}, x_{n-2}..., x_{n-r}, h_n) \quad (1)$$

In this paper we consider using as local conditioning acoustic features in order to use WaveNet as statistical vocoder (Fig. 1). WaveNet is implemented as a stack of residual blocks, where each block contains expert and gate one-dimensional dilated causal convolutions. The output of the expert and the gate are combined via element-wise multiplication. A block, i, computes a hidden state vector $z^{(i)}$, and then (due to the residual connections between layers) this is added to its input $x^{(i-1)}$ to generate its final output $x^{(i)}$:

$$z^{(i)} = \tanh(W_f^{(i)} * x^{(i-1)} + L_f^{(i)}) \odot \sigma(W_g^{(i)} * x^{(i-1)} + L_g^{(i)}) \quad (2)$$

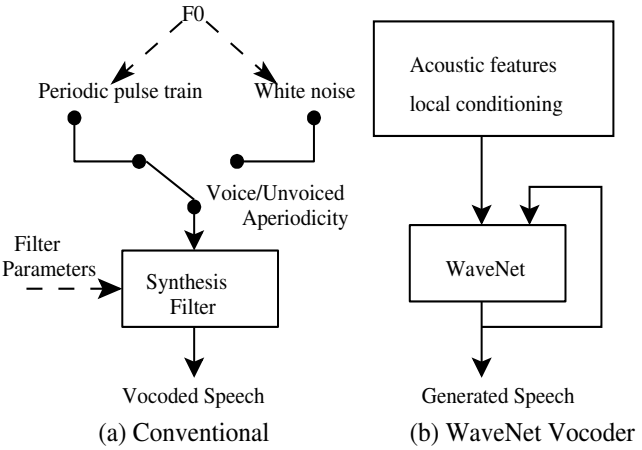

Fig. 1. Traditional vs WaveNet vocoding

In ((2)), $L_f^{(i)}$ and $L_g^{(i)}$ are the outputs for block i of the conditioning network when it is fed with h. Symbol $*$ denotes convolution and symbol $\odot$ denotes element-wise multiplication. Fig. 2 depicts the integration of acoustic features in the WaveNet architecture. The acoustic features, which are computed framewise, are of low sampling frequency (i.e., $100Hz$) and are up-sampled to the frequency of the raw waveform (i.e., $16kHz$).

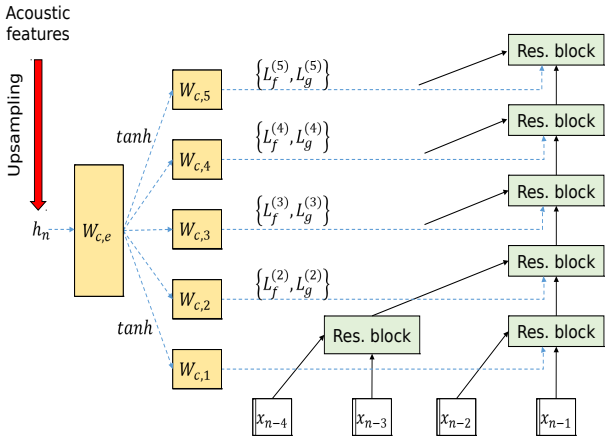

Fig. 2. Local conditioning using acoustic features

3. CONDITIONING WAVENET USING ACOUSTIC FEATURES

Speech is a non-stationary process and produced due to non-linear interaction between the various speech articulators during speech production. Conventional speech vocoders

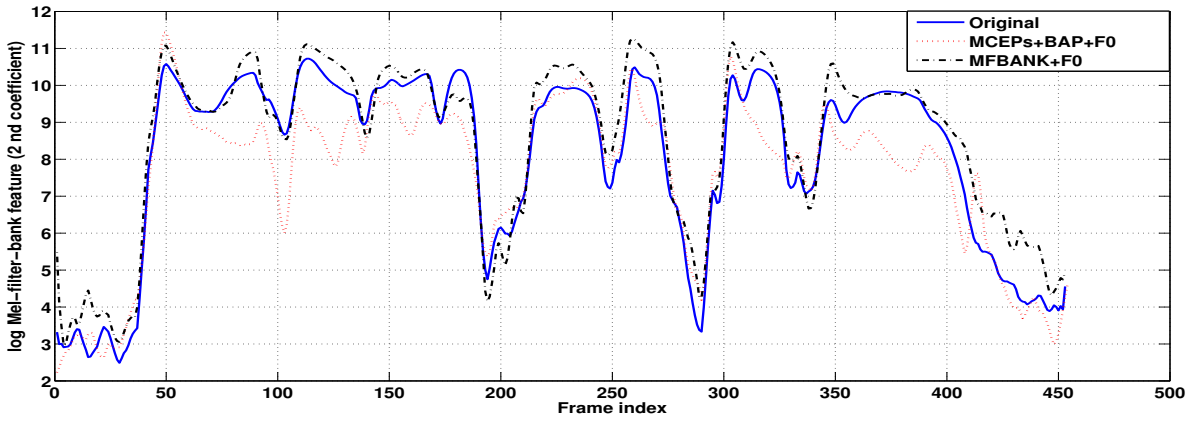

Fig. 3. 2nd log Mel-filter-bank energy computed from an original speech waveform (solid line), and from synthesized speech using MCEPs+BAP+F0 (dotted line) and MFBANK+F0 (dash-dotted line) local conditioning in the WaveNet statistical vocoder.

fail to observe these characteristics (Fig. 1(a)). In contrast, WaveNet architecture can be more accurate in representing the speech production mechanism (Fig. 1(b)). The recent work by Tamamori *et al.* [14] showed that WaveNet can indeed be used as a (statistical) vocoder by avoiding various assumption involved in a conventional speech vocoder. They have used WaveNet as vocoder by local conditioning using acoustic features like mel-cesptrum coefficients (MCEP) along with band aperiodicity (BAP) and fundamental frequency (F0) features. Note that these features are commonly used in the HMM or DNN based speech synthesis systems. MCEPs are coefficients of mel-log spectral approximation (MLSA) filter, computed from frequency warped spectrum [14].

In our work, we used mel-filterbank features (MFBANK) computed computed via Fourier Transform of the time-domain speech samples. To the limit that filter-bank bandwidths are equal to the frequency bins of the Fourier transform (magnitude spectrum) then WaveNet is expected to realize the inverse Fourier transform of the magnitude spectrum to generate the speech waveforms. Since the target however is not the minimum phase part of the speech waveform but actually the waveform itself, then WaveNet can also be seen as a mixed phase recovery mechanism. MCEPs have been used in [14] (along with BAP and F0) as acoustic features for conditioning the WaveNet. In that work, it was observed that voice of females speakers was not as well reproduced as that for male speakers. Additionally, we think that BAP features might not be necessary in case features correlated with the speech magnitude spectrum are used. Therefore, in our work we suggest to contrast MCEPs+BAP+F0 (as in [14]) with a more compact representation MFBANK+F0 with the goal to improve the overall quality of the generated speech, especially for the female speakers.

As a motivation for the discussion using MCEPs+BAP or just MFBANK, we have plotted log-filter-bank energy computed for both conditioning methods. Fig. 3 shows the comparison of the 2nd coefficient of log Mel-filter-bank energy computed from original, synthesized speech from WaveNet vocoder using MCEPs+BAP+F0 and MFBANK+F0 conditioning for one sentence taken from the ARCTIC database [18]. The contour obtained using MFBANK+F0 is closer to that from the original speech compared to that one obtained using MCEPs+BAP+F0 (as in [14]).

4. EXPERIMENTAL SETUP AND EVALUATION

We used four speakers from CMU-ARCTIC database [18]; SLT, BDL, CLB, and RMS for evaluation, where SLT and CLB are female and BDL and RMS are male. The sampling frequency is set to 16 kHz. Please note that thse are the same speakers as those used in [14]. The total number of utterances is 1,132 per speaker, and the total utterance duration is about 1 hour per speaker. We used 1050 sentences for training, 50 sentences for validation, and remaining 32 sentences were used for testing.

We have implemented a WaveNet architecture similar to the ones suggested by Deep-Voice [19] as well that one described in [14]. The differences between both the two architectures are shown in Table 1. We have found the first one faster compared to the second one for training purposes (however, no major advantages in terms of time used for waveform generation were observed). After conditioned both WaveNet architectures with the acoustic features used in [14] (MCEPs+BAP+F0), we have not noticed perceptual differences in the generated by the two architectures speech waveforms. Since the first architecture was faster in training, we used that one for all the subsequent experiments described below.

Acoustic features were extracted every $5ms$ after applying

a window of $25ms$. We have used continuous F0 and voicing decision computed from the STRAIGHT analysis [5]. MCEPs (60 coefficients) and 1 BAP feature were computed using Merlin toolkit [20]. Using standard short-time Fourier magnitude spectrum, we computed 40 MFBANK features. All these features were up-sampled to the same sampling rate of the speech ($16kHz$).

Table 1. Details of WaveNet architectures

Architecture	Proposed	Tamamori et al. [14]
Dilation Layers	50	30
Residual channels	64	256
Skip channels	256	2048
Training time	13 hr	15 hr

To find the amount of data required to train WaveNet vocoder, we have varied the number of sentences in training from 80 to 1050 sentences (80, 160, 320, 640, and 1050 sentences) only for SLT speaker. Data variation experiments are evaluated using two types of objective measure: a) short-time objective intelligibility (STOI) [21], and b) perceptual evaluation of speech quality (PESQ) [22]. STOI and PESQ results for SLT speaker with MFBANK+F0 conditioning is shown in Table 2. We have found that even with 80 sentences WaveNet vocoder gives intelligible speech but with noisy quality. Using more than 320 sentences, noise is considerably reduced. As expected, more data helps WaveNet to build a better model, which is reflected in the observed increase in STOI and PESQ scores. Note in some of the synthesized files click sounds were present. In order to remove that, acoustic features are smoothed using overlap-add method with a window size of 25 ms with a shift of 5 ms. examples of synthesized waveforms from our experiments on local conditioning as well as on data size variations can be found in [1].

Table 2. Data size variation experiments

Sentences	80	160	320	640	1050
STOI	0.64±0.04	0.67±0.05	0.72±0.04	0.78±0.06	0.86±0.03
PESQ	1.34±0.13	1.35±0.11	1.44±0.12	1.48±0.08	1.66±0.16

To know the effectiveness of acoustic features as local conditioning on WaveNet, both STOI and PESQ objective measures are computed for all testing (32 sentences) data from the 4 speakers mentioned above. Objective results are shown in Table 3. We see that MFBANK+F0 improve both STOI and PESQ when compared to the features used in [14] (MCEPs+BAP+F0).

We also evaluated the sound quality of the synthesized speech using a comparative mean opinion score (CMOS). The subjects rated sound quality of the synthesized speech using a 5-point scale: "5" for perfect, "4" for very good, "3" for good, "2" for bad, and "1" for very bad. Eighteen subjects participated in the listening experiment. The number of evaluation sentences for each subject was 40. Fig. 4 indicates the results of the CMOS test for sound quality. The error bar represents 95% confidence interval. From the figure, we first observe that indeed a more compact conditioning of the WaveNet using MFBANK+F0, as suggested in this work, provides significantly higher quality scores compared to MCEPs+BAP+F0 as it was used in [14]. Moreover, we observe that the sound quality in our case is about the same for both male and female speakers. This is not the case of samples produced using MCEPs+BAP+F0. This might be attributed to the fact that MFBANK features better represent the acoustic features for both males and females compared to MCEPs (+BAP).

[1] http://www.csd.uoc.gr/~nagaraj/

Table 3. Acoustic features local conditioning experiments

Condition	(a) STOI: Intelligibility test			
	SLT	BDL	CLB	RMS
MCEPs+BAP+F0	0.74±0.07	0.65±0.03	0.61±0.06	0.71±0.02
MFBANK+F0	0.86±0.03	0.81±0.03	0.85±0.04	0.88±0.02
	(b) PESQ: Speech quality test			
MCEPs+BAP+F0	1.34±0.11	1.35±0.17	1.33±0.11	1.37±0.13
MFBANK+F0	1.66±0.16	1.44±0.05	1.48±0.05	1.61±0.12

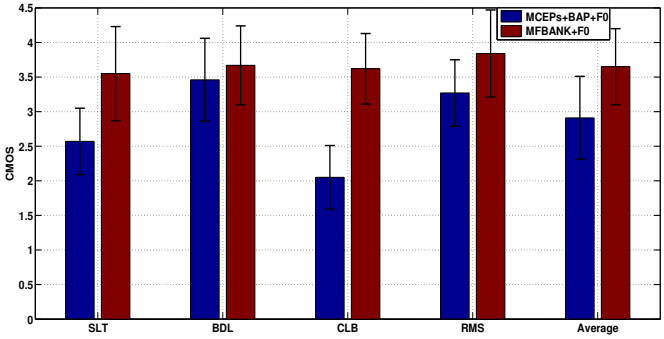

Fig. 4. Subjective listening test (CMOS).

5. CONCLUSION

In this paper, we have explored the WaveNet architecture as a speaker dependent statistical vocoder by using acoustic features as local conditioning. In that case, only 1 hour of training data are enough for producing very good quality of speech. We showed that filter-bank features are providing better local conditioning than cepstrum coefficients, allowing to produce good quality of speech for both male and female speakers. Our work was supported from objective metrics of intelligibility and sound quality as well as subjective listening tests. Future works include modeling vocoder using multi-speaker data towards effective voice conversion.

6. REFERENCES

[1] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, April 1975.

[2] Robert J McAulay and Thomas F Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 34, no. 4, pp. 744–754, 1986.

[3] Daniel Griffin and Jae Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 32, no. 2, pp. 236–243, 1984.

[4] Yannis Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 1, pp. 21–29, 2001.

[5] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain de Cheveign, "Restructuring speech representations using a pitch-adaptive time frequency smoothing and an instantaneous-frequency-based F0 extraction," *Speech Commun.*, vol. 27, pp. 187–207, 1999.

[6] Nagaraj Adiga and S R M Prasanna, "Source modeling for HMM based speech synthesis using integrated LP residual," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2016.

[7] Yannis Stylianou, Olivier Cappé, and Eric Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. speech audio process.*, vol. 6, no. 2, pp. 131–142, 1998.

[8] Tomoki Toda, Alan W Black, and Keiichi Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, 2007.

[9] A J Hunt and A W Black, "Unit selection in a concatenative speech synthesis system using a large speech database," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 1, pp. 373–376, 1996.

[10] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proc. IEEE*, vol. 101-5, pp. 1234–1252, 2013.

[11] Z. H. Ling, S. Y. Kang, H. Zen, A. Senior, M. Schuster, X. J. Qian, H. M. Meng, and L. Deng, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *IEEE Signal Process. Magazine*, vol. 32, no. 3, pp. 35–52, May 2015.

[12] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous, "Tacotron: A Fully End-to-End Text-To-Speech Synthesis Model," *arXiv preprint arXiv:1703.10135*, 2017.

[13] 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 preprint arXiv:1609.03499*, 2016.

[14] Akira Tamamori, Tomoki Hayashi, Kazuhiro Kobayashi, Kazuya Takeda, and Tomoki Toda, "Speaker-Dependent Wavenet Vocoder," in *Proc. Interspeech*, 2017, pp. 1118–1122.

[15] Kazuhiro Kobayashi, Tomoki Hayashi, Akira Tamamori, and Tomoki Toda, "Statistical voice conversion with Wavenet-Based waveform generation," in *Proc. Interspeech*, 2017, pp. 1138–1142.

[16] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, Dinei Florêncio, and Mark Hasegawa-Johnson, "Speech enhancement using bayesian WaveNet," in *Proc. Interspeech*, 2017, pp. 2013–2017.

[17] Dario Rethage, Jordi Pons, and Xavier Serra, "A WaveNet for speech denoising," *arXiv preprint arXiv:1706.07162*, 2017.

[18] John Kominek and Alan W Black, "The CMU ARCTIC speech databases," in *Proc. 5th ISCA Speech Synthesis Workshop*, 2004, pp. 223–224.

[19] Sercan Arik, Gregory Diamos, Andrew Gibiansky, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," *arXiv preprint arXiv:1705.08947*, 2017.

[20] Zhizheng Wu, Oliver Watts, and Simon King, "Merlin: An open source neural network speech synthesis system," in *Proc. SSW, Sunnyvale, USA*, 2016.

[21] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans.Audio Speech Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[22] Milos Cernak and Milan Rusko, "An evaluation of synthetic speech using the PESQ measure," in *Proc. European Congress on Acoust.*, 2005, pp. 2725–2728.