Computer Science Department
University of Crete

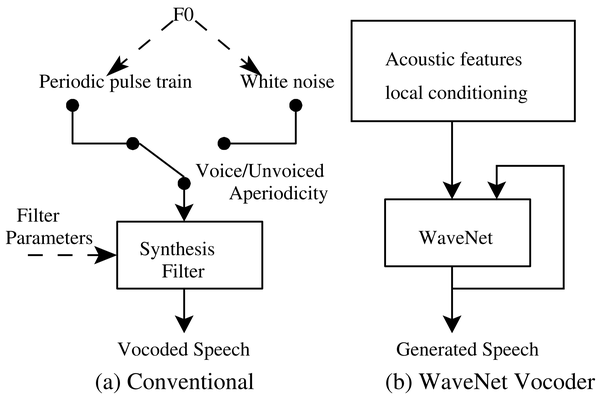
On the use of WaveNet as a Statistical Vocoder

Nagaraj Adiga, Vassilis Tsiaras, Yannis Stylianou
Speech Signal Processing Laboratory, Computer Science Department,
University of Crete, Greece
email: {nagaraj, tsiaras, yannis}@csd.uoc.gr

Introduction

- We explore the possibility of using the WaveNet architecture as a statistical vocoder [1].
- To accelerate the speech training procedure, we consider a modified version of the WaveNet as it was used in [2].
- We have showed the choice of acoustic features as local conditioning affects the quality of the generated by the WaveNet.
- Investigated the impact amount of data available for training.

Vocoder: Traditional vs WaveNet



(a) Conventional (b) WaveNet Vocoder

WaveNet Architecture

- The basic WaveNet produces babbling noise. In order to convey verbal and prosodic information in Text-to-Speech, the WaveNet is locally conditioned on linguistic and prosodic features [1, 2].
- 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.

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

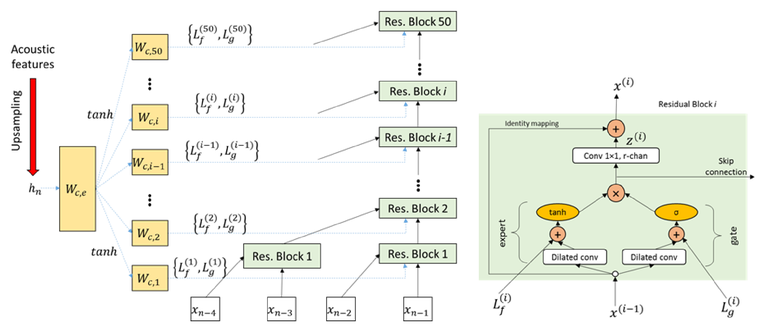
- A block, i, computes a hidden state vector $z^{(i)}$, and then added to its input $x^{(i-1)}$ to generate its final output $x^{(i)}$:

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

- In (2), $L_f^{(i)}$ and $L_g^{(i)}$ are the outputs for residual block i of the conditioning network when it is fed with h. Symbol $*$ denotes convolution and symbol $\odot$ denotes element-wise multiplication.

Conditioning WaveNet using acoustic features

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., $16kHz$). $W_{c,e}$ represents weights of conditional embedding.



Details of WaveNet architectures used in our experiments

Architecture	Proposed	Tamamori et al. [2]
Residual block (Dilation layers)	50	30
Residual channels	64	256
Skip channels	256	2048
Training time (1050 sentences)	13 hr	15 hr

Objective evaluation

- We used four speakers from CMU-ARCTIC database; SLT, BDL, CLB, and RMS for evaluation. In each speaker 1050 sentences used for training, 50 sentences for validation, and 32 sentences for testing.

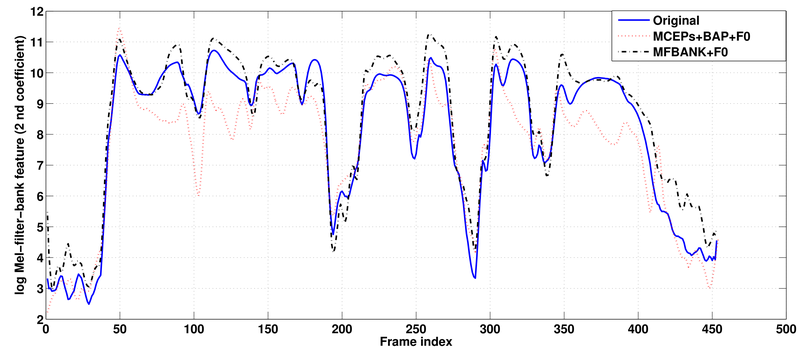
Data size variation (Number of sentences, condition: MFBANK+F0)

Method	Number of 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

Acoustic features local conditioning experiments using 1050 sentences for training

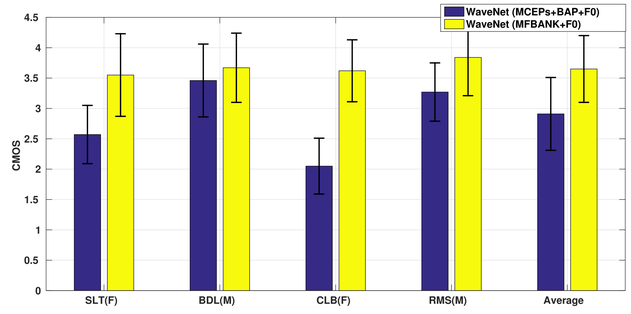
Condition	(a) STOI: Intelligibility test			
	SLT(F)	BDL(M)	CLB(F)	RMS(M)
MCEPs+BAP+F0	0.74±0.07	0.65±0.03	0.61±0.06	0.71±0.02
MFBANK+F0	0.86±0.03	0.81±0.03	0.85±0.04	0.88±0.02
	(b) PESQ: Speech quality test			
MCEPs+BAP+F0	1.34±0.11	1.35±0.17	1.33±0.11	1.37±0.13
MFBANK+F0	1.66±0.16	1.44±0.05	1.48±0.05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

Plot of 2nd log Mel-filter-bank energy WaveNet statistical vocoder



Subjective Evaluation

- Eighteen subjects participated in the listening experiment. The number of evaluation sentences for each subject was 40.



Audio files can be listened from webpage by scanning this QR code:

Conclusions and Future Work

- Explored the WaveNet architecture as a speaker dependent statistical vocoder by using acoustic features as local conditioning.
- Only 1 hour of training data are enough for producing very good quality of speech.
- Filter-bank features are providing better local conditioning than cepstrum coefficients for both Male and Female speakers.
- Future work will focus on using WaveNet vocoder for Non-parallel voice conversion.

*

References

[1] 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.

[2] Akira Tamamori, Tomoki Hayashi, Kazuhiro Kobayashi, Kazuya Takeda, and Tomoki Toda. Speaker-Dependent Wavenet Vocoder. In *Proc. Interspeech*, pages 1118–1122, 2017.