SPEECH AUGMENTATION USING WAVENET IN SPEECH RECOGNITION

Jisung Wang (jisungeeda@vuno.co) , Sangki Kim, Yeha Lee

VUNO Inc. (www.vuno.co)

(Paper ID: 4788)

Abstract

Data augmentation is crucial to improving the performance of deep neural networks by helping the model avoid overfitting and improve its generalization. In automatic speech recognition, previous work proposed several approaches to augment data by performing speed perturbation or spectral transformation. Since data augmented in these manners has similar acoustic representations with the original data, it has limited advantage in improving generalization of the acoustic model. In order to avoid generating data with limited diversity, we propose a voice conversion approach using a generative model (WaveNet), which generates a new utterance by transforming an utterance to a given target voice. Our method synthesizes speech with diverse pitch patterns by minimizing the use of acoustic features. With the Wall Street Journal dataset, we verify that our method led to better generalization compared to other data augmentation techniques such as speed perturbation and WORLD-based voice conversion. In addition, when combined with the speed perturbation technique, the two methods complement each other to further improve performance of the acoustic model.

Voice Conversion (VC) - WaveNet

WaveNet [1]

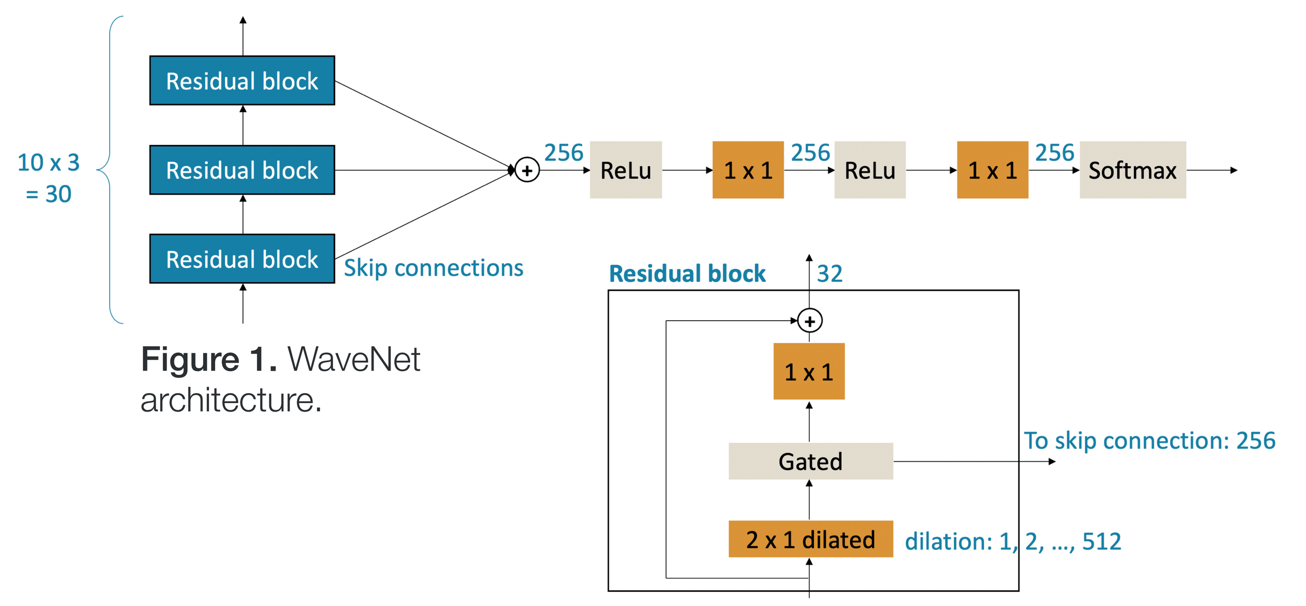

Figure 1. WaveNet architecture.

We control characteristics of generated audio by providing local and global conditions into a gated activation ftn as follows:

$$z = \tanh\left(W_{f,l} * x + V_{f,l} * y + U_{f,l} h\right) \odot \sigma\left(W_{g,l} * x + V_{g,l} * y + U_{g,l} h\right)$$

W, V, U : 2 x 1 dilated conv filter, 1 x 1 conv filter, linear projection
x : input speech waveform
y : local feature with the same time resolution as the input speech waveform
h : global feature repeated across all time steps (speaker embedding in this research)
f, g, l : filter, gate, layer index

VC-WaveNet

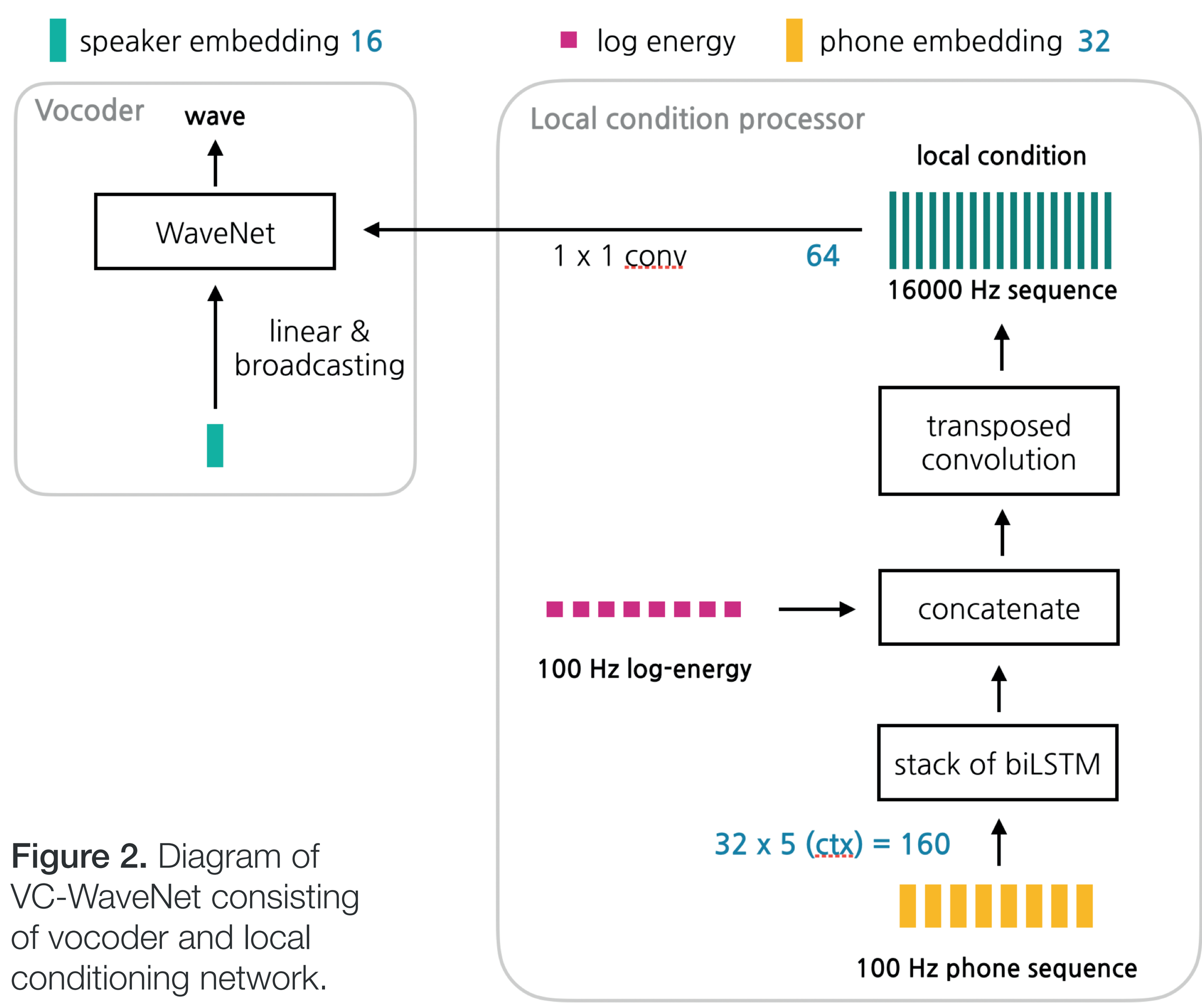

Figure 2. Diagram of VC-WaveNet consisting of vocoder and local conditioning network.

Experimental Setup

VC-WaveNet

- 2-fold training set: original, generated one
- Steps:

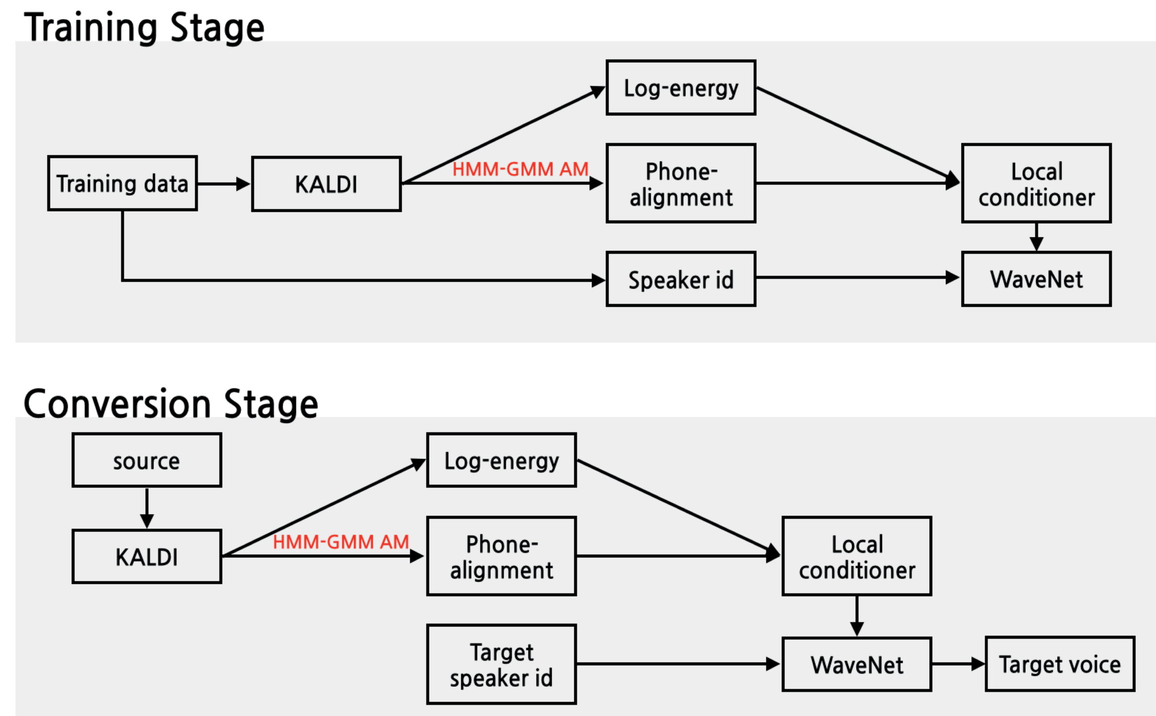

Figure 3. Training and Conversion processes of VC-WaveNet

VC-WORLD

- 2-fold training set: original, generated one
- WORLD[2] tool
- Steps:
  - Normalize log fundamental frequency F_0

$$F_0' = \frac{\sigma^y}{\sigma^x}(F_0 - \mu_x) + \mu_y$$

μ, σ : global mean and stdev of F_0 (specific speaker)
x, y : source and target spkr respectively

  - Synthesize audio with normalized parameters

Speed Perturbation

- 3-fold training set: 90%,100% (original speed), 110 %[3]
- Sox[4] audio manipulation tool

AM

- 4-layer bi-directional LSTM RNNs of 256 memory blocks
- ross entropy loss w/ SGD update
- Database: WSJ (81 hrs)

Results

AM Performance

System	Fold	Epochs	eval92 WER (%)
Baseline	1	24 / 30	5.17
Speed-perturbed	3	7 / 10	4.71
VC-WORLD	2	9 / 15	4.75
VC-WaveNet	2	12 / 15	**4.64**
VC-WaveNet + Speed-perturbed	6	5 / 5	**4.32**

Table 1. WER (%) of baseline and augmentation systems on eval92 evaluation set.

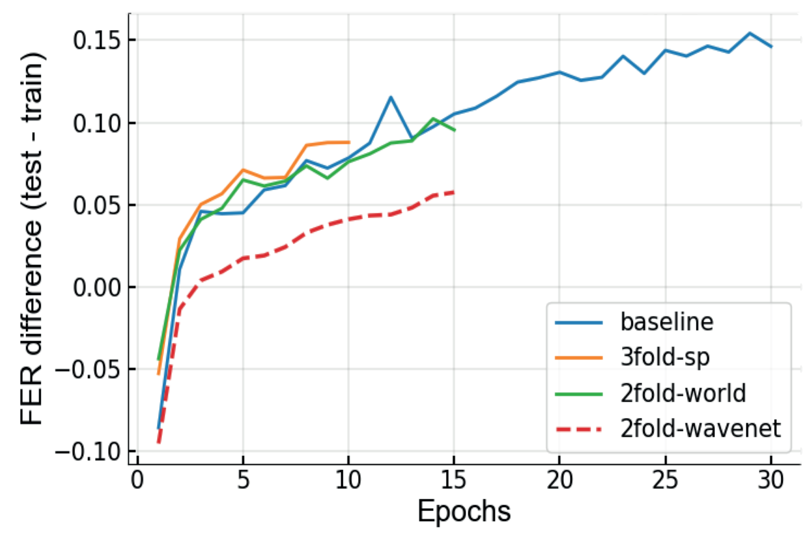

Figure 4. Difference of FER btw training and test sets across epochs. Baseline, 3-fold speed perturbation, 2-fold VC-WORLD, and 2-fold VC-WaveNet systems are compared.

Results

Novel Features

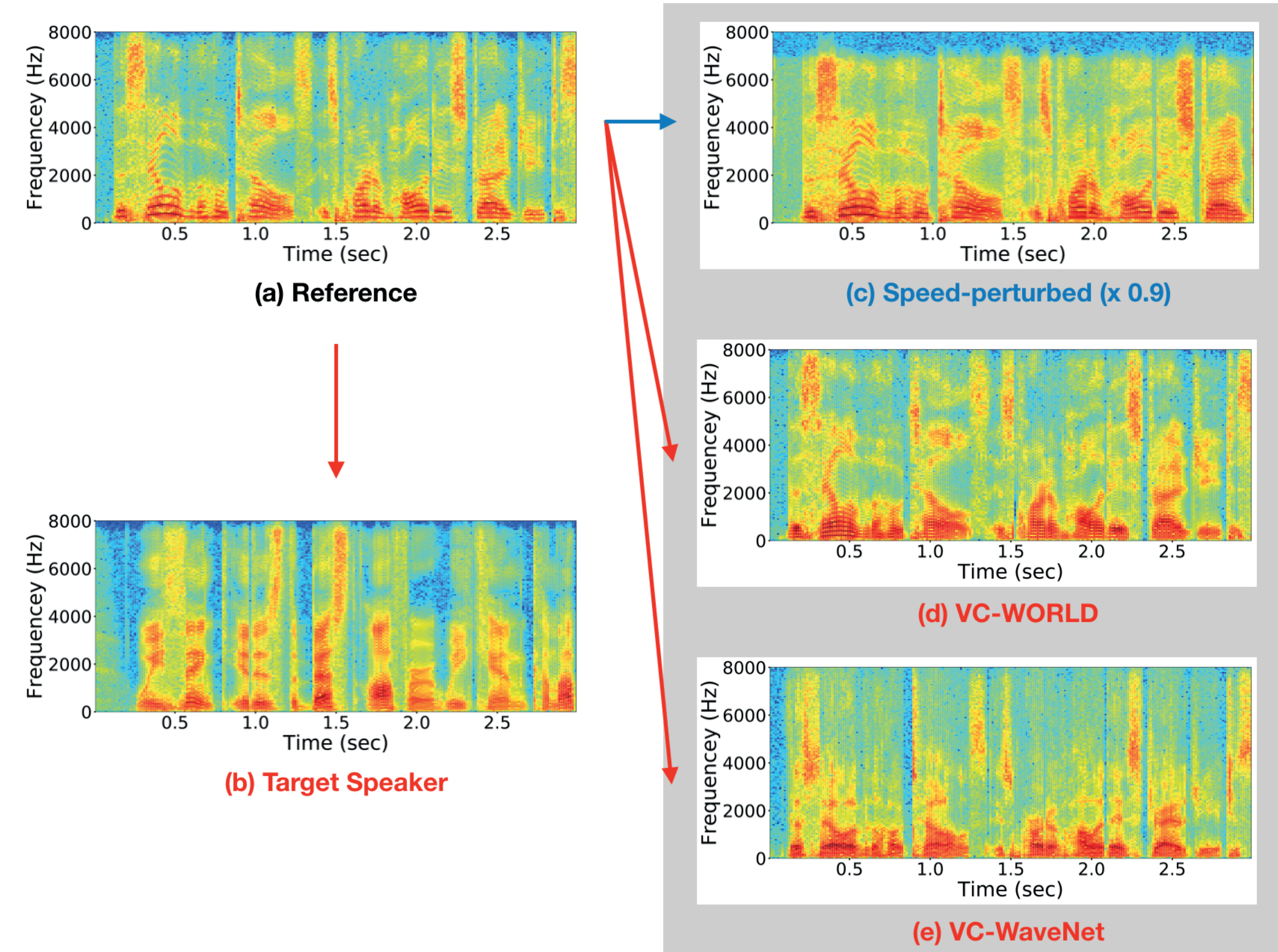

Figure 5. Spectrograms of segmented sample speech.
Ones in gray box including (c), (d) and (e) represent converted speech from the reference (a) (same linguistic content).
(a) Reference (original) utterance by source speaker '011' (female) that is to be transformed.
(b) Another utterance (different linguistic content with the reference) by target spkr '20c'(male).
(c) Speech with 90 % of original speed.
(d) Speech converted to the target voice '20c' by WORLD.
(e) Speech converted to the target voice '20c' by WaveNet.

Energy Value as a Local Condition

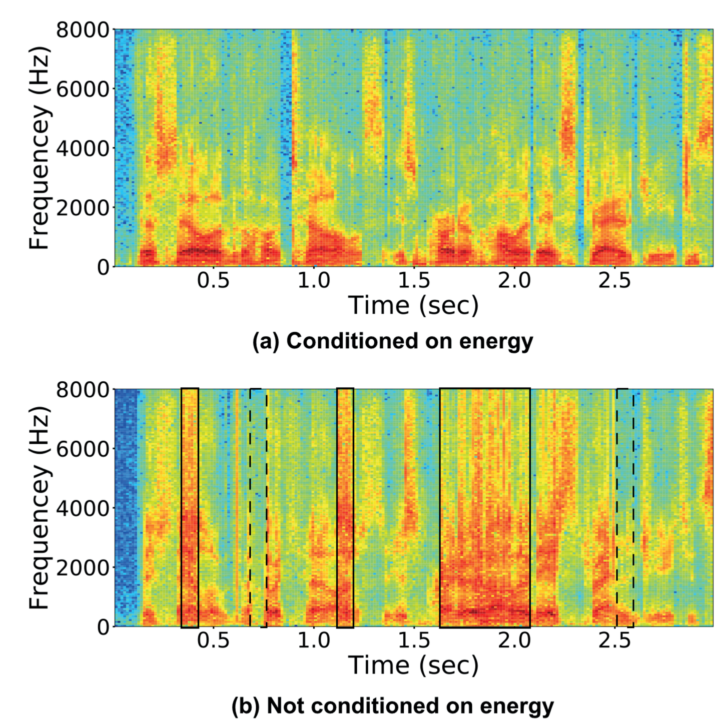

Figure 6. Spectrograms of WaveNet-based generated samples with different local condition settings.
(a) Conditioned on both linguistic feature and log-energy values.
(b) Conditioned only on linguistic feature.
Boxes with black lines denote high energy while ones with dotted lines denote low energy.

Reference

[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:1609.03499, 2016
[2] M. Morise, F. Yokomori, and K. Ozawa, "World: A vocoder-based high-quality speech synthesis system for real-time applications", IEICE Trans. Inf. Syst., vol. E99-D, no. 7m pp. 1877-1884, 2016
[3] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition", Interspeech, 2015
[4] "Sox, audio manipulation tool", Available: http://sox.sourceforge.net"