SPEECH AUGMENTATION USING WAVENET IN SPEECH RECOGNITION

Jisung Wang, Sangki Kim, Yeha Lee

VUNO Inc.
507, Gangnam-daero, Secho-gu, Seoul

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 this manner has similar acoustic representations as 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.

Index Terms— speech recognition, data augmentation, voice conversion, wavenet

1. INTRODUCTION

As the capacity of deep neural networks (DNNs) increases, large training dataset with rich patterns becomes more important. However, it is expensive and time consuming to build labeled dataset. One common strategy to deal with this problem is data augmentation (DA), which increases the quantity of training data by adding transformed samples that preserve the original labels. The most common DA approaches in automatic speech recognition (ASR) include speed perturbation [1] and vocal tract length perturbation (VTLP) [2]. However, the speech data generated from these methods have limited diversity as their acoustic features resemble the original ones.

Thus, we propose a novel method using a generative model to increase diversity in synthetic data. Specifically we propose VC-WaveNet, a voice conversion (VC) technique using WaveNet [3] as a generative model. VC refers to transforming an original utterance into a new utterance that resembles the voice of a target speaker while preserving the linguistic content.

Our approach has the following contributions. First, our technique employs a generative model. Previous techniques [2], [4] have focused on feature modifications. The downside of feature modification is that the acoustic model (AM) might recognize the relation between the original and the transformed data. Once the AM learns the relationship, the augmented data is no longer a new representation and its benefit will be limited.

Second, our technique generates utterances with WaveNet instead of conventional vocoders. Conventional vocoders lose detail information during parameterization, which causes artifacts to be left in the generated speech [5]. These artifacts, which do not exist in real data, can hinder the AM's generalization on real data.

Lastly, our technique generates speech with diverse pitch patterns. Our WaveNet is not conditioned by vocoder parameters such as a fundamental frequency and spectral information. Consequently, as details of acoustic features including pitch variability is not given to WaveNet, the synthetic speech can have diverse pitch patterns.

We present results on Wall Street Journal (WSJ) corpus to show that the proposed technique performs better than DA using speed perturbation which is the most effective DA so far. In addition, VC-WaveNet combined with speed perturbation showed further improvement by mutually complementing each other.

2. RELATION TO PRIOR WORK

With the improvements in deep generative models, training synthetic data is widely performed for various tasks [6], [7], [8]. In speech recognition, however, most studies have proposed techniques mainly about modification of existing acoustic data such as speed perturbation [1] and VTLP [2].

Recently Nishizaki [9] proposed a DA method using a variational autoencoder (VAE) for the first time and showed it improves an AM. We also introduce a DA with generative model but with WaveNet, which is more suitable for synthesizing waveform data.

WaveNet, firstly proposed as a part of TTS system with linguistic conditioning features [3], [10], [11], has been re-

cently investigated as a vocoder with acoustic features for various purposes including voice conversion [12], [13], [14], [15], [16]. Unlike these studies, our final goal is to augment data for AM training. For this purpose we convert voice with diverse pitch patterns by omitting detailed acoustic information and we show the synthetic speech in this manner actually improves ASR performance.

3. WAVENET

Wavenet, a generative model proposed in [3] for producing high quality signal, is an autoregressive network which directly estimates a raw waveform sample-by-sample. For the input sequence $x = x_1, ..., x_N$, the model approximates the joint probabilities of signal as follows:

$$p(x) \simeq \prod_{n=1}^{N} p(x_n | x_{n-R-1}, x_{n-R}, ..., x_{n-1}, \Lambda), \quad (1)$$

where R represents a receptive field length and Λ represents model parameters. Its architecture mainly consists of several stacks of residual blocks including 2 $\times$ 1 dilated causal convolution, gated activation, and 1 x 1 convolution. The gated activation function in a residual block is defined as:

$$z = tanh(W_{f,l} * x) \odot \sigma(W_{g,l} * x), \quad (2)$$

where $*$ denotes a causal convolution operator, $\odot$ denotes an element-wise product operator, l is the layer index, f and g denote a filter and a gate, respectively, and W is a trainable 2 $\times$ 1 convolution filter. Every skip connections from residual blocks are led to separate 1 $\times$ 1 convolution layers and finally softmax layer. Then the softmax layer generates the posterior probabilities of waveform quantized to 256 values by μ-law compressor[17].

We can control characteristics of generated audio by providing conditions [3], [10], [18]. The conditional WaveNet has the gated activation function in the form of the following:

$$z = tanh(W_{f,l} * x + V_{f,l} * y + U_{f,l} h) \quad \odot \\ \sigma(W_{g,l} * x + V_{g,l} * y + U_{g,l} h), \quad (3)$$

where V is a learnable 1 $\times$ 1 convolution filter, U is a learnable linear projection, y represents processed local auxiliary features which have the same time resolution as the input speech waveform, and h is global features which is repeated across time.

The architecture of WaveNet model used in this study is shown in Fig. 1. This model consists of the local conditioning network and the waveform generator. The conditioning network first encodes linguistic features with 3-layer bi-directional LSTM RNNs and concatenates them with another auxiliary local feature, log-energy values. Note that both local features are extracted every 10 msec from the 25 msec length

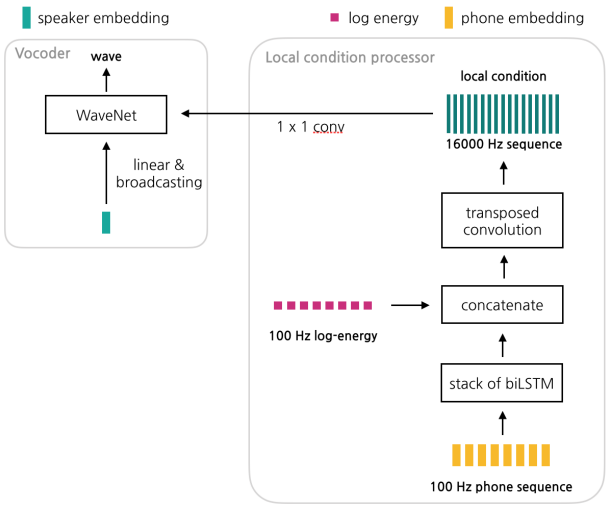

Fig. 1. Diagram of a variant WaveNet consisting of vocoder and local conditioning network.

of windowed speech waveform, resulting in 100 Hz of sampling frequency. Thus it upsamples to the desired frequency (16 kHz) with a transposed convolution layer, which is followed by 1-D convolution. Finally the waveform synthesis part generates audio at 16 kHz conditioned on outputs from conditioning network and speaker embedding vector based on Eq. 2.

4. EXPERIMENTAL SETUP

Our experiments are done on the WSJ corpus (LDC93S6B and LDC94S13B) [19], containing 16 kHz, 16 bit of reading speech. Its training set contains 283 speakers, and 10 speakers for "dev93" validation set, 8 speakers for "eval92" test set. Each speaker has around 20 to 30 minutes of audio files.

4.1. ASR

For training DNN AM, we used 4-layer bi-directional LSTM RNNs of 256 memory blocks with cross entropy loss. It was trained with forced-alignments of triphones obtained from previously trained GMM-HMM AM. Its input feature is a filter-bank with 40 coefficients on a mel-scale plus energy value and their first and second temporal derivatives, which leads to 123 dimensional input vector per one frame. For filter-bank feature, we used 25 msec long hamming window every 10 msec and each input feature was normalized for each individual utterance. During decoding, a beam-size of 10 was used with a pruned trigram language model.

Model parameters are updated by stochastic gradient descent (SGD) on all cases using a fixed learning rate of 1e-3 and a mini-batch size of 16. The number of training epochs used was 30 for the baseline system while a lower number of

epochs was used for the augmented systems to keep its training time similar to the baseline system. We used "dev 93" dataset as a validation set for early stopping. Results are presented on the "eval 92" dataset.

4.2. VC-WaveNet

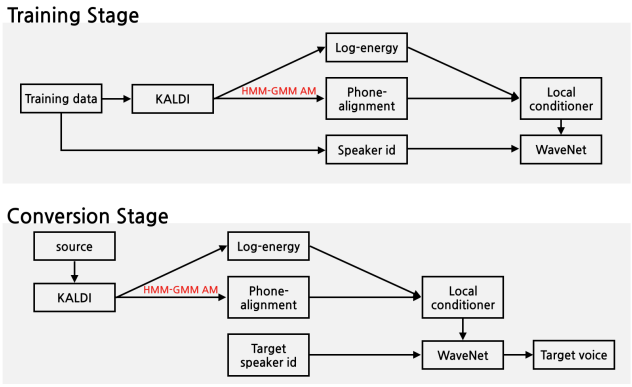

Fig. 2. Training and Conversion processes of VC-WaveNet.

Fig. 2 illustrates the VC system with the conditional WaveNet presented in Sec. 3. In order to train the VC system, we first trained a GMM-HMM system on the training set with Kaldi [20] recipe s5. The GMM-HMM AM had 3392 triphones, upon which we extracted forced-alignments to prepare data for DNN AM. Then the alignment was converted into phoneme sequences which were then used to train the WaveNet model. As input into the LSTM layers, we used phoneme context data which include two previous and two following phonemes for each phoneme. Log-energy values, another local auxiliary feature, was also obtained in the form of 1 dimensional sequence during the process. Finally we set speaker information as the global condition in the form of a trainable embedding vector.

During conversion (Fig. 2), we provided one random speaker id out of total 283 speakers for each reference utterance. This process led to 2-fold data augmented system.

We used nv-wavenet source code provided from NVIDIA to speed up inference time. Training set of 80 hours was split into 4 and each was assigned one GPU device. As a result, 4 GPUs each parallelly processed 20 hours of training data. In this manner, we were able to complete the conversion process in 12 hours.

Hyper-parameters of WaveNet model is as follows: 30 layers, 64 residual channels, 256 skip channels, and 512 as a maximum dilation length. Input sequence representing linguistic info are set as 160 (32 embedding dimension $\times$ 5 context width) dimensional vector sequence. Also we set a speaker embedding dimension as 16. Dropout was applied with probability of 0.15 for all 3 LSTM layers in a conditioning network and on linear transformation layer on speaker embedding.

4.3. VC-WORLD

We also made another 2-fold system with VC but using conventional vocoder called WORLD [21]. With its analysis algorithms, WORLD estimates speech parameters including F_0, spectral envelope, and aperiodic values. For VC, we adopted a simple method of globally normalizing the source speakers mean and standard deviation of the log fundamental frequency F_0 in frame-level by using a linear transformation:

$$F_0{}' = \frac{\sigma^y}{\sigma^x}(F_0 - \mu_x) + \mu_y, \tag{4}$$

where μ and σ are global mean and standard deviation values of log F_0, x and y are source and target speaker in training data respectively.

4.4. Speed Perturbation

Speed-perturbation technique is reported to be the most effective augmentation method for various sizes of corpora [1]. Following [1], two speed-perturbed copies of the original training data were generated by changing the speed to 90 % and 110 % of the original speed using Sox audio manipulation tool [22]. For this 3-fold training set, we generated alignments using GMM-HMM system using Kaldi [20].

5. RESULTS AND DISCUSSION

Table 1 shows the results on "eval92" test set with different data augmented systems. A relative improvement of 10.3 % was observed using WaveNet-based converted training data while 8.9 % was obtained when using speed perturbed data. WORLD-based synthetic data is found to be less helpful even than speed perturbed data.

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

A sample of augmented data with 90 % of original speed is shown in Fig. 3 (c). A shift in the signal power towards lower frequencies and resulting low power at high frequencies are observed as expected. Also we can see expanded spectrogram in the time axis as a result of time warping. Its spectral shape is similar to the reference in Fig. 3 (a).

Fig. 3 (d) and (e) describe voice audio samples converted from Fig. 3 (a) to the target voice Fig. 3 (b). For synthetic

speech from WORLD, its feature shape seems similar to the reference, just as the speed-perturbed one does. On the other hand, the spectrogram of speech generated by WaveNet differs from spectral of the reference although it contains same linguistic contents. This synthetic utterance (Fig. 3 (e)) represents a new generated utterance, novel as if spoken by different speaker rather than a modified one.

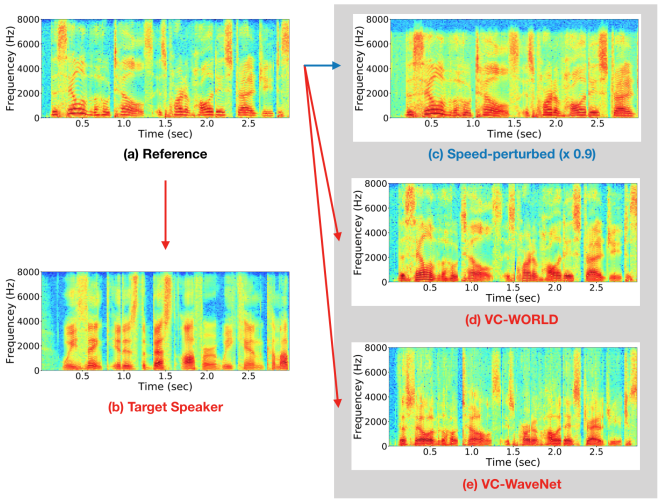

Fig. 3. Spectrograms of segmented sample speech. Ones in gray box including (c), (d) and (e) represents converted speech from the reference (a) (same linguistic content).
(a) Reference (original) utterance by source speaker 011 (female) that is to be transformed. (b) Another utterance (different linguistic content with the reference) by target speaker 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.

According to Fig. 4, augmented data in this manner seems to improve generalization as it reduces the gap of frame error rate (FER) between training and validation sets compared to other augmented data systems.

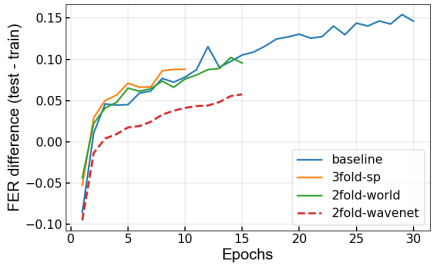

Fig. 4. Difference of FER between training and test sets across epochs. Baseline, 3-fold speed perturbation, 2-fold VC-WORLD, and 2-fold VC-WaveNet systems are compared.

Speed perturbation technique has an advantage in that it is an effective DA method with a low implementation cost [1]. Also it warps signals in the time axis which is not done in VC-WaveNet. Combining these two complementary methods achieved the lowest WER of 4.32 % as shown in Table 1.

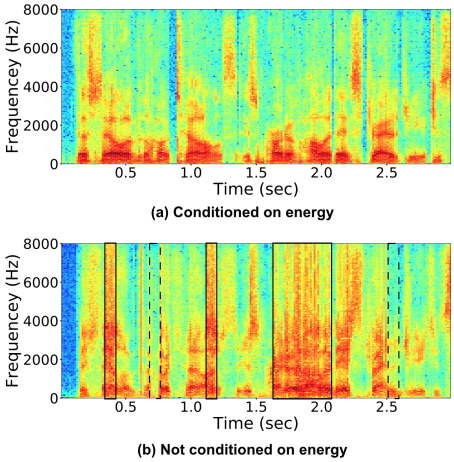

Fig. 5. Spectrograms of WaveNet-based generated samples with different local condition settings.
(a) Conditioned on both linguistic feature and log-energy values. (b) Conditioned only on linguistic feature. Boxes with black lines denote high signal energy while ones with dotted lines denote low signal energy.

According to [3], F_0 information as a local condition for WaveNet is crucial to produce speech with natural prosody. In order to generate utterances with various stress and pitch changes across sentence, we did not condition on F_0 values. However, conditioning only on linguistic feature without any acoustic feature caused generating an unstable waveform with uncontrolled volume (Fig. 5) which is also difficult for human to recognize. Thus we additionally provided log-energy values as local conditions which led to stable synthetic waveform.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a VC-WaveNet technique without vocoder parameters and demonstrate its effectiveness as a DA method for ASR task. As our WaveNet is locally conditioned only on linguistic and energy information, it generates speech imitating the target speaker with diversity unlike other VC cases. It is shown to improve WER of WSJ corpus better than speed perturbation, an most successful existing DA method, or a simple VC by WORLD vocoder. Furthermore, combining VC-WaveNet and speed perturbation led to better WER. In future we would like to investigate the perturbation of local conditions to the WaveNet to further improve ASR performance. Also we intend to study the effectiveness of our proposed techniques when applied to a larger corpus.

7. REFERENCES

[1] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," *Interspeech*, 2015.

[2] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," *ICML Workshop on Deep Learning for Audio, Speech, and Language Processing*, 2013.

[3] 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.

[4] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural netowkr acoustic modeling," *ICASSP*, pp. 100–104, 2014.

[5] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for wavenet vocoder," *Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 712–718, 2017.

[6] Z. Wang, J. Yang, H. Jin, E. Shechtman, A. Agarwala, J. Brandt, and T. Huang, "Deepfont: Identify your font from an image," *Proceedings of ACMM*, 2015.

[7] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," *arXiv:161207828*, 2016.

[8] J. Liu, H. Rahmani, N. Akhtar, and A. Mian, "Learning human pose models from synthesized data for robust rgb-d action recognition," *arXiv:1707.00823v2*, 2018.

[9] H. Nishizaki, "Data augmentation and feature extraction using variational autoencoder for acoustic modeling," *Proceedings of APSIPA Annual Summit and Conference*, 2017.

[10] S. O. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman, s. Sengupta, and M. Shoeybi, "Deep voice: Real-time neural text-to-speech," *ICLR*, 2017.

[11] S. O. Arik, G. Diamos, A. Gibiansky, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice2: Multi-speaker neural text-to-speech," *NIPS*, 2017.

[12] N. Adiga, V. Tsiaras, and Y. Stylianou, "On the use of wavenet as a statistical vocoder," *ICASSP*, 2018.

[13] K. Chen, B. Chen, J. Lai, and K. Yu, "High-quality voice conversion using spectrogram-based wavenet vocoder," *Interspeech*, 2018.

[14] K. Kobayashi, T. Hayashi, A. Tamamori, and T. Toda, "Statistical voice conversion with wavenet-based waveform generation," *Interspeech*, 2017.

[15] K. Kobayashi, T. Hayashi, A. Tamamori, and T. Toda, "Statistical voice conversion with wavenet-based waveform generation," *Interspeech*, 2017.

[16] L. J. Liu, Z. H. Ling, Y. Jiang, M. Zhou, and L. R. Dai, "Wavenet vocoder with limited training data for voice conversion," *Interspeech*, 2018.

[17] ITU-T, "Recommendation g. 711," *Pulse Code Modulation (PCM) of voice frequencies*, 1988.

[18] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent wavenet vocoder," *Interspeech*, 2017.

[19] D. B. Paul and J. M. Baker, "The design for the wall street journal-based csr corpus," *Proceedings of the workshop on Speech and Natural Language*, pp. 357–362, 1992.

[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

[21] 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. 7, pp. 1877–1884, 2016.

[22] "Sox, audio manipulation tool," *Available: http://sox.sourceforge.net/*.