Phonetic Posterior Grams에 의해 조건화된 적대적 생성 신경망을 사용한 음성 변환 시스템

임진수 · 강천성 · 김동하 · 김경섭*

충남대학교

Voice Conversion using Generative Adversarial Nets conditioned by Phonetic Posterior Grams

Jin-su Lim · Cheon-seong Kang · Dong-Ha Kim · Kyung-sup Kim*

Chungnam National University

E-mail : dla0830ld@cnu.ac.kr / kcs93023@cnu.ac.kr / kdongha1103@cnu.ac.kr / sclkim@cnu.ac.kr

요 약

본 논문은 매핑 되지 않은 입력 음성과 목표음성 사이에 음성 변환하는 비 병렬 음성 변환 네트워크를 제안한다. 기존 음성 변환 연구에서는 변환 전후 스펙트로그램의 거리 오차를 최소화하는 방법을 주로 학습 한다. 이러한 방법은 MSE의 이미지를 평균 내는 특징으로 인하여 생성된 스펙트로그램의 해상도가 저하되는 문제점이 있었다. 또한, 병렬 데이터를 사용해 연구를 진행했기 때문에 데이터를 수집하는 것에도 어려움이 많았다. 본 논문에서는 입력 음성의 발음 PPGs를 사용하여 비 병렬 데이터 간 학습을 진행 하며, GAN 학습을 통해 더욱 선명한 음성을 생성하는 방법을 사용하였다. 제안한 방법의 유효성을 검증하기 위해서 기존 음성 변환 시스템에서 많이 사용하는 GMM 기반 모델과 MOS 테스트를 진행하였으며 기존 모델에 비하여 성능이 향상되는 결과를 얻었다.

ABSTRACT

This paper suggests non-parallel-voice-conversion network conversing voice between unmapped voice pair as source voice and target voice. Conventional voice conversion researches used learning methods that minimize spectrogram's distance error. Not only these researches have some problem that is lost spectrogram resolution by methods averaging pixels. But also have used parallel data that is hard to collect. This research uses PPGs that is input voice's phonetic data and a GAN learning method to generate more clear voices. To evaluate the suggested method, we conduct MOS test with GMM based Model. We found that the performance is improved compared to the conventional methods.

키워드

Generative Adversarial Network, cGAN, Voice Conversion, Phonetic Posterior Grams, Tacotron

Ⅰ. 서 론

음성 변환은 한 화자에 음성을 다른 화자에 음성 특성에 맞추어 변환하는 것을 말한다. 기존의 음성변환 연구에서는 다른 화자가 같은 문장을 말하는 Pair된 병렬데이터를 이용하여 Gaussian Mixture Models(GMM) 기반에 음성 변환을 하는 [1]과 [2], Bidirectional Long Short-Term Memory 기반의 [3]등이 있으며 Pair된 데이터 없이 PPGs(Phonetic Posterior Grams)를 중간에 생성하여 단계적으로 음성 변환하는 [4]등이 있다.

대부분의 음성 변환 연구에서는 변환 음성에 대한 스펙트로그램을 생성하고 실제 스펙트로그램과 오차 평균인 Mean squared error(MSE)에 기반하여 학습을 한다. 하지만 MSE를 사용한 학습은 생성된 스펙트로그램 이미지와 정답을 평균하려는

* corresponding author

성향이 강하기 때문에 생성되는 결과에 해상도가 떨어지는 문제가 발생한다. 이러한 문제를 생성 모델에서 좋은 성능을 내고 있는 GAN(Generative Adversarial Networks) [5] 구조를 사용해 해결하고자 한다. 또한, PPGs를 이용하여 입력 음성의 발음을 인식하는 단계와 TTS(Text-To-Speech) 분야의 대표적인 모델인 Tacotron[6]의 음성 합성 모듈을 사용해 성능을 개선한 비 병렬 데이터 간 음성 변환 모델을 제안한다.

본 논문의 2장에서는 해당 연구 이전에 연구되었던 사전 연구에 대한 내용을 다룰 것이며, 3장에서는 제안된 모델의 전반적인 설명을 다룰 것이다. 마지막으로 4장에서는 제안된 모델과 기존의 음성 변환 시스템과의 성능 차이를 다룬다.

Ⅱ. 사전 연구

cGAN(Conditional Generative Adversarial Nets)[7]은 임의의 잡음 z로부터 생성된 데이터를 조절하기 위하여 모델 입력에 조건을 추가한 모델이다. 추가적인 정보를 제공해 판별기는 생성 데이터와 조건이 매칭 되는지를 판단하며 생성기에서 생성되는 데이터를 조절할 수 있다.

PPGs(Phonetic Posterior Grams)는 특정 시간 간격에 대한 발음 종류들의 사후 확률을 나타낸 행렬이다. 이러한 발음 종류들은 단어, 음소 등의 단위로 표현되고 나뉠 수 있다. [4]에서 제안된 PPGs 기반 음성 합성 시스템은 Kaldi speech toolkit[8]을 사용하여 구현된 모델을 통하여 PPGs을 생성한다. 이를 이용하여 Pair된 데이터 없이 목표음성만으로 DBLSTM(deep bidirectional long short-term memory) 모델을 학습한다.

본 논문에서는 비 병렬 데이터 간 음성 변환을 위하여 SI-ASR을 신경망으로 구현하여 SR-Model (Speech Recognition Model)로써 사용하였다.

Tacotron은 구글에서 발표한 end-to-end TTS 모델이다. Tacotron은 입력 문자열을 스펙트로그램으로 출력하는 RNN Attention 메커니즘 기반의 인코더와 디코더로 구성이 되어 있다.

CBHG 모듈은 Tacotron에서 사용한 신경망 블록이다. Conv1D Bank, Highway Net, Bi-GRU로 구성되어 있다. 입력 데이터에 부분적인 특징을 잡아내는데 Conv filer를 사용하며, 이를 RNN을 이용하여 연속적으로 데이터를 처리하여 결과를 만들어 내는 모듈이다. 본 논문에서는 Tacotron의 디코더를 기반을 두어 GAN 모델의 생성기를 구성하였다.

Ⅲ. 방법

3.1 제안된 모델

제안된 모델은 기존 PPGs를 이용한 음성변환 모델[4]을 기반으로 GAN의 학습 모델을 설정하였다[그림 1]. 제안된 모델의 구성 단계는 순차적으로 SR-Model 학습, GAN Model 학습 그리고 음성변환 과 같이 3단계로 진행 된다.

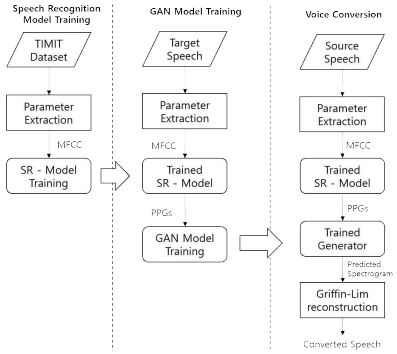

그림 1. 제안된 모델의 전체 구조

3.1.1 SR-Model 구조

SR-Model[그림 2]은 시간단위로 각 음소별 확률을 나타낸 PPGs를 출력하도록 만들어진 모델로 [4]에서 사용한 SI-ASR 모델을 CBHG 모듈과 FC Layer를 이용하여 구현하였다. 입력 음원에 대하여 MFCC 특징을 추출하여 모델의 입력으로 사용하였으며 생성한 출력을 각 음소 클래스에 대한 분류 학습을 진행하였다. 다음 단계에 GAN 모델 학습 때, SR-Model을 학습 하지 않도록 하여 SR-Model에서 출력된 PPGs를 생성기 학습 중 고정적으로 출력되도록 하였다.

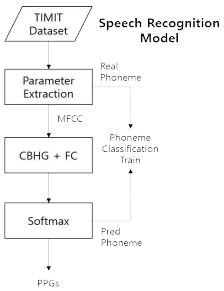

그림 2. SR 모델의 구조

3.1.2 GAN 모델 구조

생성기와 판별기로 적대적 학습을 하는 구조로 구성 및 학습을 하여 변환 단계에서 학습된 생성기를 이용하였다.

생성기 모델의 입력으로 학습된 SR-Model을 이용하여 생성된 목표음성 PPGs를 입력 잡음 z와 붙여 사용하였다. 또한 입력 잡음과 붙이지 않은 PPGs를 Attention 메커니즘의 memory로서 사용하였다.

생성기 구조로는 Tacotron의 Decoder의 구조를 사용하여 구성을 하였다. prenet과 Attention RNN, GRU[9], CBHG 모듈을 사용하였으며, 이를

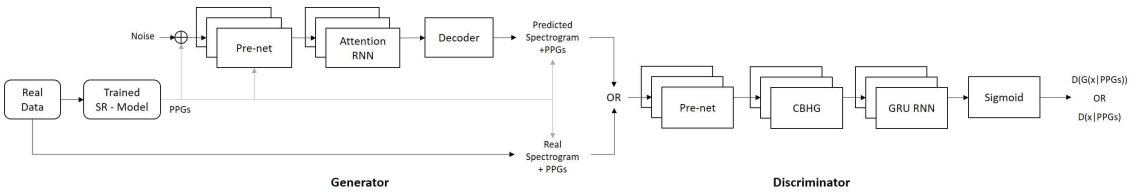

그림 3. GAN 모델 구조

통하여 목표음성의 Linear-Scale 스펙트로그램을 생성하는 네크워크를 구성하였다.

판별기 구조는 Prenet 모듈, CBHG 모듈, GRU, Sigmoid를 이용하여 구성되었으며 입력으로 넣어준 데이터에 대하여 참인지 거짓인지 판별하였다.

판별기 모델의 입력으로는 생성기에서 생성된 이미지와 생성기의 입력 PPGs를 묶어 거짓 데이터를 생성하며, 실제 데이터를 SR-Model에 입력하여 생성한 PPGs와 묶어 참 데이터를 만들었다.

최종적인 생성기의 학습을 위한 GAN 모델은 [그림 3]과 같이 구성하였다.

3.1.3 음성 변환

음성 변환은 입력음성의 MFCC를 SR-Model에 입력하여 생성된 PPGs를 학습된 생성기를 거쳐 목표음성의 스펙트로그램으로 변조한다.

변환된 스펙트로그램을 Vocoder[10]를 이용하여 음성으로 재구성하여 음성을 생성하였다.

3.2 제안된 모델의 손실 함수

제안된 모델의 손실 함수는 conditional GAN에서 사용한 방식에 Reconstruction loss를 추가하여 구성을 하였다. 기본적인 Adversarial loss는 식 (1)과 같으며 판별기는 식을 최대화하는 방향으로 생성기는 최소화하는 방향으로 학습을 한다.

$$L_{cGAN}(G,D) = E_{x\sim P_{data}(x)}[\log D(x|PPGs)] + E_{z\sim P_z(z)}[\log(1-D(G(z|PPGs))) \quad (1)$$

생성기에서 판별기를 속이기 위한 역할뿐만 아니라 기존에 Ground truth와 유사한 스펙트로그램을 생성하기 위하여 pix2pix[11]에서 사용한 방법인 L_1 거리를 추가적으로 생성기에서 loss function 으로 이용하였다. 생성기에서 예측된 스펙트로그램과 실제 입력 음성에 스펙트로그램 간 L_1 거리를 생성기의 GAN loss인 식 (1)에 식 (2)을 추가하였다.

$$L_{L_1}(G) = E_{x\sim P_{data}(x)}[\|x - G(z|ppgs)\|_1] \quad (2)$$

최종적으로 모델의 손실함수는 식 (3)과 (4)와 같다

$$D = \arg max_D L_{cGAN}(G,D) \quad (3)$$

$$G = \arg min_G (L_{cGAN}(G,D) + L_{L_1}(G)) \quad (4)$$

3.3 Baseline 모델

GAN을 이용한 모델과에 성능을 비교하기 위해서 Baseline 모델과 비교를 하였으며 오픈소스인 FestVox system[12]의 Voice Conversion Toolkit을 기반으로 Baseline을 구성하였다.

Baseline 모델은 GMM 모델을 기반으로 만들어졌다. 훈련 시에는 가우시안 혼합의 수를 64로 설정 하였으며, 추가 리소스 없이 훈련데이터로만 훈련을 진행하였다.

Ⅳ. 실험

4.1 실험 환경

본 연구에서는 음성 인식 모델에 학습을 위하여 TIMIT Corpus[13]을 사용했으며, PPGs 발음클래스로써 61개의 영어 음소를 사용하였다. 음성 인식 모델의 음소 분류 정확도는 53%의 정확도를 가지고 진행하였다.

GAN 구조의 음성 합성 모델 학습 단계에서 ARCTIC Corpus[14]을 사용을 하여 모델을 학습하였다. 음성 평가를 위해서 ARCTIC Corpus에서 일부를 나누어 훈련데이터와 검증 데이터로 사용했다.

모델 학습 중 모델 가중치에 대한 기울기의 발산을 막기 위해 Gradient Clipping [15]을 사용하였으며 적대적인 예들에 대한 네트워크의 취약성을 줄이기 위해 positive라벨의 값을 0.9로 하는 One sided label smoothing[16]을 사용하였다.

각 모델에 학습 횟수는 baseline은 프로그램에 기본 설정 값을 이용하였으며 제안된 모델은 생성기에 loss가 수렴하여 더 이상 변화하지 않을 때까지 진행하였다. Adam Optimize을 사용했고, learning rate 값을 3e-4로 설정하였으며 약 790 epoch에 학습을 진행하였다.

하드웨어 사양은 Intel i5 8400 2.8GHZ, NVIDIA GTX 1080을 사용하여 SR-Model에서는 10시간 소요 되었으며 baseline은 8시간, GAN 모델은 30시간 소요 되었다.

4.2 실험 평가

각 모델별 생성된 음성을 비교하기 위해서 MOS(Mean Opinion Score) 테스트를 평가방법으로 진행하였다. MOS는 생성된 음성에 자연스러움과 발음의 명확함 정도를 1점에서 5점으로 평가하여 실시하였다. 테스트에는 정상 청력을 가진 남여

17명을 대상으로 실시하였다. ARCTIC Corpus를 나눈 검증 데이터를 이용하였으며 다른 성별 간 전환 음성을 이용하여 진행하였다.

표 1. MOS 테스트 결과

	음질	발음
Baseline	2.18	3.59
GAN	2.65	3,53

다음은 남성에서 여성으로의 목소리 변조의 MOS 테스트 결과[표 1]를 보여준다. 음질에 대한 MOS결과는 제안된 모델이 Baseline 모델보다 0.47점 더 높은 결과를 보였다. 발음의 정확도에 대한 MOS결과는 제안된 모델이 0.06점 낮은 결과를 보였다. 제안된 모델이 Baseline모델보다 음질 면에서 큰 성능 향상을 보여주었고, 발음에서는 비슷한 성능을 보여주었다.

Ⅴ. 결 론

본 논문에서는 한 화자에 음성을 다른 화자에 음성 특성에 맞추어 변환하는 음성 변환을 PPGs와 GAN 구조를 이용하여 수행하였다. 실험결과에서 제안된 모델이 Baseline 모델에 비하여 생성 음성에 대해 MOS에서 개선된 결과를 보여주었다. 그러나 SR-Model의 분류 정확도에 따라 음성 변환에 성능 저하가 관측되었다. 향후 연구로는 raw한 데이터를 입력하여 특징 추출부터 합성까지 전체를 모델로 학습하는 end-to-end 방식을 이용하여 SR-Model에서 생기는 정보 손실을 줄일 수 있다면 더욱 선명하고 자연스러운 음성 변환이 가능할 것으로 기대한다.

Acknowledgement

본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 SW중심대학사업의 연구결과로 수행되었음(2015-0-00930)

References

[1] Stylianou, Yannis, Olivier Cappe, and Eric Moulines, "Continuous probabilistic transform for voice conversion.", IEEE Transactions on speech and audio processing 6.2, 131-142, 1998.

[2] Toda, Tomoki, Alan W. Black, and Keiichi Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory.", IEEE Transactions on Audio, Speech, and Language Processing 15.8, 2222-2235, 2007.

[3] L. Sun, S. Kang, K. Li, and H. Meng, “Voice conversion using deep bidirectional Long Short-Term Memory based Recurrent Neural Networks,” in Proc. ICASSP, 2015.

[4] Sun, Lifa, et al. "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training.", 2016 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2016.

[5] Goodfellow, Ian, et al., "Generative adversarial nets.", Advances in neural information processing systems. 2014.

[6] Wang, Yuxuan, et al., "Tacotron: A fully end-to-end text-to-speech synthesis model.", arXiv preprint, 2017.

[7] Mirza, Mehdi, and Simon Osindero, "Conditional generative adversarial nets.", arXiv preprint arXiv:1411.1784, 2014.

[8] Povey, Daniel, et al., "The Kaldi speech recognition toolkit.", IEEE 2011 workshop on automatic speech recognition and understanding. No. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[9] Chung, Junyoung, et al., "Empirical evaluation of gated recurrent neural networks on sequence modeling.", arXiv preprint arXiv:1412.3555, 2014.

[10]Griffin, Daniel, and Jae Lim, "Signal estimation from modified short-time Fourier transform.", IEEE Transactions on Acoustics, Speech, and Signal Processing 32.2, 236-243, 1984.

[11] Isola, Phillip, et al., "Image-to-image translation with conditional adversarial networks.", arXiv preprint ,2017.

[12] Anumanchipalli, Gopala Krishna, Kishore Prahallad, and Alan W. Black., "Festvox: Tools for creation and analyses of large speech corpora.", Workshop on Very Large Scale Phonetics Research, UPenn, Philadelphia, 2011.

[13] Garofolo, John S., "TIMIT acoustic phonetic continuous speech corpus.", Linguistic Data Consortium, 1993.

[14] Kominek, John, and Alan W. Black., "The CMU Arctic speech databases.", Fifth ISCA workshop on speech synthesis, 2004.

[15] Pascanu, Razvan, Tomas Mikolov, and Yoshua Bengio., "On the difficulty of training recurrent neural networks.", International Conference on Machine Learning, 2013.

[16] Salimans, Tim, et al., "Improved techniques for training gans.", Advances in Neural Information Processing Systems, 2016.