(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0084443
(43) 공개일자 2020년07월13일

(51) 국제특허분류(Int. Cl.)
G10L 13/033 (2013.01) *G06N 3/08* (2006.01)
G10L 21/013 (2013.01) *G10L 25/18* (2013.01)
G10L 25/30 (2013.01)
(52) CPC특허분류
G10L 13/033 (2013.01)
G06N 3/08 (2013.01)
(21) 출원번호 10-2018-0169788
(22) 출원일자 2018년12월26일
심사청구일자 2018년12월26일

(71) 출원인
충남대학교산학협력단
대전광역시 유성구 대학로 99 (궁동, 충남대학교)
(72) 발명자
김경섭
서울특별시 강남구 일원로14길 25 108동 1402호
강천성
대전광역시 서구 대덕대로 18, 402호
(뒷면에 계속)
(74) 대리인
특허법인 공간

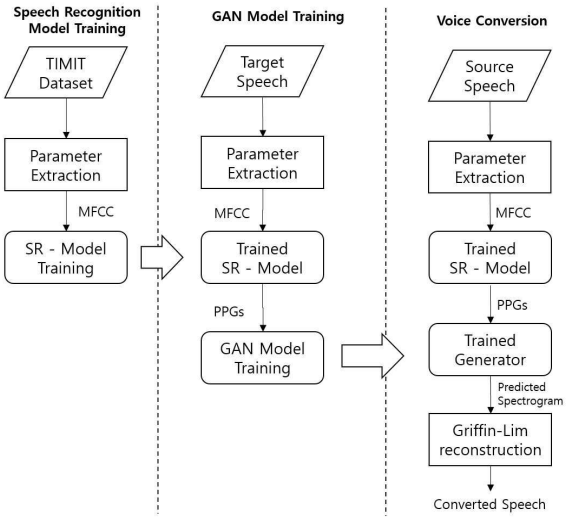
전체 청구항 수 : 총 2 항

(54) 발명의 명칭 **음성 변조 시스템 및 방법**

(57) 요 약

본 발명에 의한 음성 변조 시스템은, SR 모델 학습 모듈, GAN 모델 학습 모듈, 음성 변조 모듈을 포함하는 음성 변조 시스템으로서, 상기 SR 모델 학습 모듈은, 입력 음원에 대해 MFCC 특징을 추출하여 SR 모델 학습을 진행하고, GAN 모델 학습 모듈은, 목표 음성을 SR 모델에 입력하여 PPGs 를 생성한 후 GAN 모델 학습을 진행하고, 음성 변조 모듈은, 입력음성의 MFCC를 SR 모델에 입력하여 PPGs를 생성하고 GAN 구조를 통해 학습된 생성기를 거쳐 목표 음성의 스펙트로그램으로 변조한 후, 이를 음성으로 재구성하여 음성을 생성하는 것을 특징으로 한다.

대 표 도 - 도1



(52) CPC특허분류
G10L 25/18 (2013.01)
G10L 25/30 (2013.01)
G10L 2021/0135 (2013.01)
(72) 발명자
김동하
대전광역시 유성구 자운로157번길 132, 914동 302호

임진수
울산광역시 북구 염포로 599 현대자동차사택 105동 601호

이 발명을 지원한 국가연구개발사업
과제고유번호 2015-0-00930
부처명 과학기술정보통신부
연구관리전문기관 정보통신기술진흥센터
연구사업명 SW전문인력양성/정보통신창의인재양성
연구과제명 2018년 SW중심대학 지원사업_충남대
기 여 율 1/1
주관기관 충남대학교 산학협력단
연구기간 2018.01.01 ~ 2018.12.31

명 세 서

청구범위

청구항 1

SR 모델 학습 모듈, GAN 모델 학습 모듈, 음성 변조 모듈을 포함하는 음성 변조 시스템으로서,

상기 SR 모델 학습 모듈은, 입력 음원에 대해 MFCC 특징을 추출하여 SR 모델 학습을 진행하고,

GAN 모델 학습 모듈은, 목표 음성을 SR 모델에 입력하여 PPGs 를 생성한 후 GAN 모델 학습을 진행하고,

음성 변조 모듈은, 입력음성의 MFCC를 SR 모델에 입력하여 PPGs를 생성하고 GAN 구조를 통해 학습된 생성기를 거쳐 목표 음성의 스펙트로그램으로 변조한 후, 이를 음성으로 재구성하여 음성을 생성하는 것을 특징으로 하는 음성 변조 시스템.

청구항 2

SR 모델 학습 모듈, GAN 모델 학습 모듈, 음성 변조 모듈을 포함하는 음성 변조 시스템을 이용한 음성 변조 방법으로서,

상기 SR 모델 학습 모듈이, 입력 음원에 대해 MFCC 특징을 추출하여 SR 모델 학습을 진행하는 제1 단계;

상기 GAN 모델 학습 모듈이, 목표 음성을 SR 모델에 입력하여 PPGs 를 생성한 후 GAN 모델 학습을 진행하는 제2 단계;

상기 음성 변조 모듈이, 입력음성의 MFCC를 SR 모델에 입력하여 PPGs를 생성하고 GAN 구조를 통해 학습된 생성기를 거쳐 목표 음성의 스펙트로그램으로 변조한 후, 이를 음성으로 재구성하여 음성을 생성하는 제3 단계;

를 포함하는 것을 특징으로 하는 음성 변조 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 음성 변조 시스템 및 방법에 관한 것으로, 보다 상세하게는 한 화자에 음성을 다른 화자에 음성 특성에 맞추어 변환하는 음성 변조 시스템 및 방법에 관한 것이다.

배 경 기 술

[0002] 음성 변조(음성 변환)은 한 화자에 음성을 다른 화자에 음성 특성에 맞추어 변환하는 것을 말한다.

[0003] 기존의 음성변환 기술로는, 비특허문헌 1과 비특허문헌 2에 기재된 것(다른 화자가 같은 문장을 말하는 Pair된 병렬데이터를 이용하여 Gaussian Mixture Models(GMM) 기반에 음성 변환을 하는 것), 비특허문헌 3에 기재된 것(Bidirectional Long Short-Term Memory 기반의 음성 변환 기술) 등이 있으며, 비특허문헌 4에 기재된 것과 같이, Pair된 데이터 없이 PPGs(Phonetic Posterior Grams)를 중간에 생성하여 단계적으로 음성 변환하는 기술 등이 있다.

[0004] 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 성향이 강하기 때문에 생성되는 결과에 해상도(음질)가 떨어지는 문제가 발생한다.

선행기술문헌

비특허문헌

[0006] (비특허문헌 0001) 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.

(비특허문헌 0002) 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.

(비특허문헌 0003) 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.

(비특허문헌 0004) 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.

발명의 내용

해결하려는 과제

[0007] 본 발명이 해결하고자는 하는 과제는, 변조 후 음질이 향상되는 음성 변조 시스템 및 방법을 제공하는 것이다.

과제의 해결 수단

[0009] 본 발명에 의한 음성 변조 시스템은, SR 모델 학습 모듈, GAN 모델 학습 모듈, 음성 변조 모듈을 포함하는 음성 변조 시스템으로서, 상기 SR 모델 학습 모듈은, 입력 음원에 대해 MFCC 특징을 추출하여 SR 모델 학습을 진행하고, GAN 모델 학습 모듈은, 목표 음성을 SR 모델에 입력하여 PPGs 를 생성한 후 GAN 모델 학습을 진행하고, 음성 변조 모듈은, 입력음성의 MFCC를 SR 모델에 입력하여 PPGs를 생성하고 GAN 구조를 통해 학습된 생성기를 거쳐 목표 음성의 스펙트로그램으로 변조한 후, 이를 음성으로 재구성하여 음성을 생성하는 것을 특징으로 한다.

[0010] 본 발명에 의한 음성 변조 방법은, SR 모델 학습 모듈, GAN 모델 학습 모듈, 음성 변조 모듈을 포함하는 음성 변조 시스템을 이용한 음성 변조 방법으로서, 상기 SR 모델 학습 모듈이, 입력 음원에 대해 MFCC 특징을 추출하여 SR 모델 학습을 진행하는 제1 단계; 상기 GAN 모델 학습 모듈이, 목표 음성을 SR 모델에 입력하여 PPGs 를 생성한 후 GAN 모델 학습을 진행하는 제2 단계; 상기 음성 변조 모듈이, 입력음성의 MFCC를 SR 모델에 입력하여 PPGs를 생성하고 GAN 구조를 통해 학습된 생성기를 거쳐 목표 음성의 스펙트로그램으로 변조한 후, 이를 음성으로 재구성하여 음성을 생성하는 제3 단계; 를 포함하는 것을 특징으로 한다.

발명의 효과

[0012] 본 발명에 의한 음성 변조 시스템 및 방법은, 변조 후 음질이 향상된다.

도면의 간단한 설명

[0014] 도 1은 본 발명의 음성 변조 순서도

도 2는 본 발명의 SR 모델의 구조

도 3은 본 발명의 GAN 모델의 구조

발명을 실시하기 위한 구체적인 내용

[0015] 본 발명은 다양한 변경을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 상세하게 설명하고자 한다. 그러나 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변환, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다. 본 발명

을 설명함에 있어서 관련된 공지 기술에 대한 구체적인 설명이 본 발명의 요지를 흐릴 수 있다고 판단되는 경우 그 상세한 설명을 생략한다.

[0016] 본 출원에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다.

[0018] 대부분의 음성 변환 연구에서는 변환 음성에 대한 스펙트로그램을 생성하고 실제 스펙트로그램과 오차 평균인 MSE에 기반하여 학습을 한다. 하지만 MSE를 사용한 학습은 생성된 스펙트로그램 이미지와 정답을 평균하려는 성향이 강하기 때문에 생성되는 결과에 해상도가 떨어지는 문제가 발생한다. 이러한 문제를 생성 모델에서 좋은 성능을 내고 있는 GAN 구조[{Goodfellow, Ian, et al., "Generative adversarial nets.", Advances in neural information processing systems. 2014.} 참조]를 추가해 해결하고자 한다. 또한, PPGs를 이용하여 입력 음성의 발음을 인식하는 단계를 거치고 TTS 분야의 대표적인 모델인 Tacotron[{Wang, Yuxuan, et al., "Tacotron: A fully end-to-end text-to-speech synthesis model.", arXiv preprint, 2017. 6.} 참조]의 음성 합성 모듈을 사용해 성능을 개선한 비 병렬 데이터 간 음성 변환을 할 수 있다.

[0020] 본 발명에 의한 음성 변조 시스템 및 방법은 기존 PPGs를 이용한 음성변환 모델을 기반으로 GAN의 학습 모델을 설정하였다. 그 방법의 변조 순서도는 도 1과 같다. 도 1은 본 발명의 음성 변조 순서도이다.

[0021] 도 1의 순서도의 구성 단계는, SR 모델 학습(Speech Recognition Model Training), GAN 모델 학습(GAN Model Training), 음성 변환(음성 변조)의 3단계로 진행된다.

[0022] 이때, SR 모델 학습을 하는 모듈을 SR 모델 학습 모듈, GAN 모델 학습을 하는 모듈을 GAN 모델 학습 모듈, 음성 변환(음성 변조)를 하는 모듈을 음성 변조 모듈이라고 하면, 본 발명의 음성 변조 시스템은 SR 모델 학습 모듈, GAN 모델 학습 모듈, 음성 변조 모듈을 포함한다.

[0023] Parameter Extraction 은, 입력 음성 데이터에서 정보를 추출하여 Mel-frequency cepstral coefficients(MFCC)를 추출하는 단계를 의미한다.

[0024] SR 모델 학습은 MFCC를 입력으로하여 PPGs를 생성하는 모델 훈련 단계를 의미한다.

[0025] Griffin_Lim Reconstruction 은 그리핀 림 알고리즘을 바탕으로 음성을 생성하는 단계를 의미한다.

[0027] SR 모델 학습에 대해 설명하면 다음과 같다.

[0028] 도 2는 본 발명의 SR 모델의 구조이다.

[0029] SR 모델 학습은, 도 2에서와 같이, 시간단위로 각 음소별 확률을 나타낸 PPGs를 출력하도록 만들어진 모델로 [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.]에서 사용한 SI-ASR 모델을 Dropout[Srivastava, Nitish, et al., "Dropout: a simple way to prevent neural networks from overfitting.", The Journal of Machine Learning Research 15.1, 1929-1958, 2014.]을 적용한 2층의 FC Layer인 Prenet 모듈, Tacotron의 CBHG 모듈과 FC Layer를 이용하여 구현하였다. 입력 음원에 대하여 MFCC 특징을 추출하여 입력으로 사용하였으며 생성한 출력을 Softmax와 argmax를 통해 각 음소 클래스에 대한 분류 학습을 진행하였다. GAN 모델 학습 때에는 argmax를 하지 않고 Softmax를 통해 모든 클래스에 대한 벡터를 가지도록 하여 PPGs를 생성을 하며, SR-Model을 학습 하지 않도록 하여 생성기 학습 중 입력 음성에 PPGs를 고정적으로 출력되도록 하였다.

[0030] Phoneme Classification 은 음소를 분류하는 단계로, 단위 시간마다 들어오는 음성을 음소 클래스에 맞춰 사후 확률을 계산하여 PPGs를 생성한다.

[0032] GAN 모델 학습에 대해 설명하면 다음과 같다.

[0033] 본 발명의 GAN 모델은 음성 합성을 위한 생성기 부분을 기존의 음성합성을 위해 잘 알려진 모델인 Tacotron에

디코더를 기반으로 도 3과 같은 생성기를 구성하였다.

[0034] 도 3은 본 발명의 GAN 모델의 구조이다.

[0035] 목표 음성을 SR 모델에 입력하여 PPGs을 생성했으며 입력 잡음 z와 붙여 Prenet 모듈을 통과시킨 값을 Attention RNN [Vinyals, Oriol, et al., "Grammar as a foreign language.", Advances in Neural Information Processing Systems, 2015.] 입력으로 사용하였다. 또한 입력 잡음과 붙이지 않은 PPGs를 Attention 메커니즘의 memory로서 입력을 해주었다.

[0037] Attention RNN의 결과를 Decoder로써 Residual connection을 포함한 GRU[Chung, Junyoung, et al., "Empirical evaluation of gated recurrent neural networks on sequence modeling.", arXiv preprint arXiv:1412.3555, 2014.] 2개를 사용하였으며 Prenet과 CBHG 모듈을 통과하여 목표 음성의 Linear-Scale 스펙트로그램을 생성하였다.

[0038] 판별기 구조는 Prenet 모듈, CBHG 모듈, GRU, Sigmoid를 이용하여 구성되었으며 입력으로 넣어준 데이터에 대하여 참인지 거짓인지 판별하도록 하였다. 판별기 모델의 입력으로는 거짓 데이터와 참 데이터를 입력하도록 구성하였다. 생성기에서 생성된 이미지와 생성기의 입력 PPGs를 묶어 거짓 데이터를 생성하며, 실제 데이터를 SR-Model에 입력하여 PPGs를 생성하고 이를 묶어 참 데이터를 만들었다. 이러한 입력 데이터를 판별기에 교차로 입력하였다.

[0040] GAN(Generative Adversarial Networks)모델은, Generator(생성기, 스펙토그램 생성기)와 Discriminator(판별기, 스펙토그램 판별기)의 두 가지 모델이 적대적으로 학습을 하는 방식의 네트워크이다.

[0041] Generator는 가짜 데이터를 생성하는 모델이다. Generator의 목적은 Discriminator가 판별하지 못할 정도로 정교한 데이터를 생성하는 모델이다.

[0042] 원하는 데이터를 생성하는 모델로 제안된 모델에서는 타코트론의 구조를 가져와서 만들었다. 현 모델에서는 스펙트로그램을 생성한다. 스펙트로그램은, 소리의 스펙트럼을 시각화하여 그래프로 표현하는기법이다. 시간상 진폭축의 변화를 시각적으로 볼 수 있는 파형과 주파수상 진폭축의 변화를 시각적으로 볼 수 있는 스펙트럼의 특징이 모두 결합된 구조로 시간축과 주파수상의 진폭의 차이를 농도나 표시 색상으로나타낸다.

[0043] 생성자로부터 생성된 데이터와 실제데이터를 섞어서 Discriminator에게 보여주며 해당 모델(Discriminator)은 입력으로 들어온 데이터가 진짜인지 가짜인지 판별하는 모델이다.

[0044] GAN 구조는 이러한 Generator와 Discriminator가 서로 적대적으로 학습을 하며 점점 더 진짜같은 데이터를 생성하며 판별로 정교한 데이터를 더 잘 판별하도록 하여 기존 생성 모델에 비하여 샤프한 데이터를 생성하게된다.

[0045] 타코트론은, 구글에서만든 Text to Speech 모델로 제안된 모델의 제너레이터 부분을 구현하였으며 해당 논문에서 생성자(Generator)의 구조를 가져왔다.

[0046] CBHG 는, 타코트론에서 음성 데이터의 특징 추출을 위해 사용된 모듈로 컨볼루션레이어와 Fully connected로 구성된 Highwaynet, GRU로 구성된모듈 로 타코트론 논문에서는 High-level feature를 뽑기 위해 사용한 네트워크라고 한다.

[0047] Atenttion 은 주어진 데이터에서 어느 부분에 집중을 할 것인가를 찾아내는 부분으로 Attention RNN을 넣는 것이 모델의 Generalization(일반화)에 도움이 된다고 한다. 모델의 일반화란 “학습되지 않은 문제에 대해 얼마나 좋은 성능을 보여줄 수 있는가”를 뜻한다. RNN은 히든 노드가 방향을 가진 엣지로 연결돼 순환구조를 이루는(directed cycle) 인공신경망의 한 종류이다.

[0049] 음성 변환 과정은 입력음성의 MFCC를 SR-Model에 입력하여 PPGs를 생성하고 GAN 구조를 통해 학습된 생성기를 거쳐 목표 음성의 스펙트로그램으로 변조하였다. 이를 Griffin-Lim Vocoder를 이용하여 음성으로 재구성하여 음성을 생성하였다.

[0051] 제안된 모델의 손실 함수는 conditional GAN에서 사용한 방식에 Reconstruction loss를 추가하여 구성을 하였다. 기본적인 Adversarial loss는 다음의 식 (1)과 같으며, 판별기는 식을 최대화하는 방향으로, 생성기는 최소화하는 방향으로 학습을 한다.

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

[0052] ---(1)

[0054] 생성기에서 판별기를 속이기 위한 역할뿐만 아니라 기존에 Ground truth와 유사한 스펙트로그램을 생성하기 위하여 pix2pix[{Isola, Phillip, et al., "Image-to-image translation with conditional adversarial networks.", arXiv preprint ,2017.} 참조]에서 사용한 방법인 거리를 추가적으로 생성기에서 loss function으로 이용하였다. 생성기에서 예측된 스펙트로그램과 실제 입력 음성에 스펙트로그램 간 거리를 생성기의 GAN loss인 식 (1)에 다음의 식 (2)을 추가하였다.

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

[0055] ---(2)

[0057] 최종적으로 모델의 손실함수는 다음의 식 (3)과 (4)와 같다

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

[0059] ---(3)

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

[0060] ---(4)

[0062] GAN을 이용한 모델과에 성능을 비교하기 위해서 Baseline 모델과 비교를 하였으며 오픈소스인 FestVox system[{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.} 참조]의 Voice Conversion Toolkit을 기반으로 Baseline을 구성하였다.

[0063] Baseline 모델은 GMM 모델을 기반으로 만들어 졌다. 훈련 시에는 가우시안 혼합의 수를 64로 설정 하였으며, 추가 리소스 없이 훈련데이터로만 훈련을 진행하였다.

[0065] 본 발명의 모델에 의한 실험결과는 다음과 같다.

[0066] 본 발명에서는 음성 인식 모델에 학습을 위하여 TIMIT Corpus[{Garofolo, John S., "TIMIT acoustic phonetic continuous speech corpus.", Linguistic Data Consortium, 1993.} 참조]을 사용했으며, PPGs 발음클래스로써 61개의 영어 음소를 사용하였다. 음성 인식 모델의 음소 분류 정확도는 53%의 정확도를 가지고 진행하였다.

[0067] GAN 구조의 음성 합성 모델 학습 단계에서 ARCTIC Corpus[{Kominek, John, and Alan W. Black., "The CMU Arctic speech databases.", Fifth ISCA workshop on speech synthesis, 2004.} 참조]을 사용하여 모델을 학습하였다. 음성 평가를 위해서 ARCTIC Corpus에서 일부를 나누어 훈련데이터와 검증 데이터로 사용했다.

[0068] 모델 학습 중 모델 가중치에 대한 기울기의 발산을 막기 위해 Gradient Clipping [{Pascanu, Razvan, Tomas Mikolov, and Yoshua Bengio., "On the difficulty of training recurrent neural networks.", International

Conference on Machine Learning, 2013.} 참조]을 사용하였으며 적대적인 예들에 대한 네트워크의 취약성을 줄이기 위해 positive라벨의 값을 0.9로 하는 One sided label smoothing[{Salimans, Tim, et al., "Improved techniques for training gans.", Advances in Neural Information Processing Systems, 2016.} 참조]을 사용하였다.

[0069] 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.

[0070] 하드웨어 사양은 Intel i5 8400 2.8GHZ, NVIDIA GTX 1080을 사용하여 SR-Model에서는 10시간 소요 되었으며 baseline은 8시간, GAN 모델은 30시간 소요 되었다.

[0072] 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.

[0073] 다음의 표 1은, 남성에서 여성으로의 목소리 변조의 MOS 테스트 결과를 보여준다.

[0074] *표 1*

	음질	발음
Baseline	2.18	3.59
GAN	2.65	3,53

[0076] 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.

[0078] 따라서 본 발명에 의한 음성 변조 방법은, 다음의 3단계를 포함한다.

[0079] (1) 제1 단계: SR 모델 학습 모듈이, 입력 음원에 대해 MFCC 특징을 추출하여 SR 모델 학습을 진행하는 단계

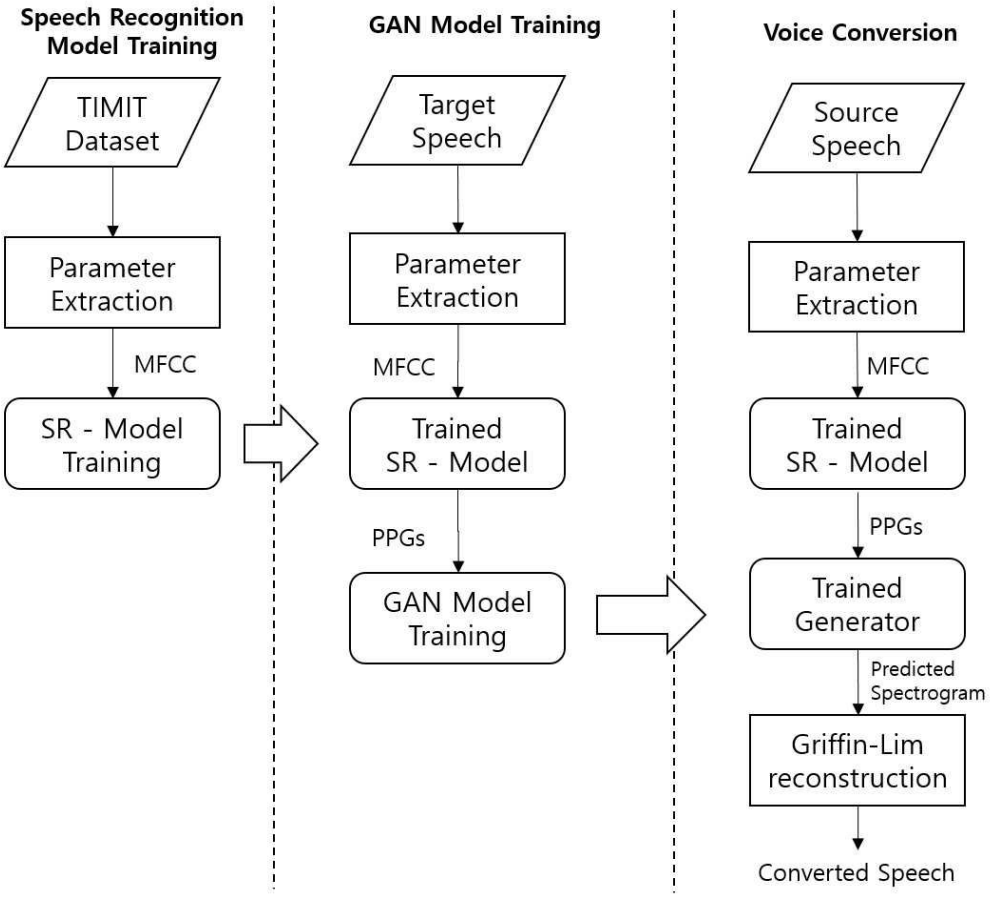
[0080] (2) 제2 단계: GAN 모델 학습 모듈이, 목표 음성을 SR 모델에 입력하여 PPGs 를 생성한 후 GAN 모델 학습을 진행하는 단계

[0081] (3) 제3 단계: 음성 변조 모듈이, 입력음성의 MFCC를 SR 모델에 입력하여 PPGs를 생성하고 GAN 구조를 통해 학습된 생성기를 거쳐 목표 음성의 스펙트로그램으로 변조한 후, 이를 음성으로 재구성하여 음성을 생성하는 단계

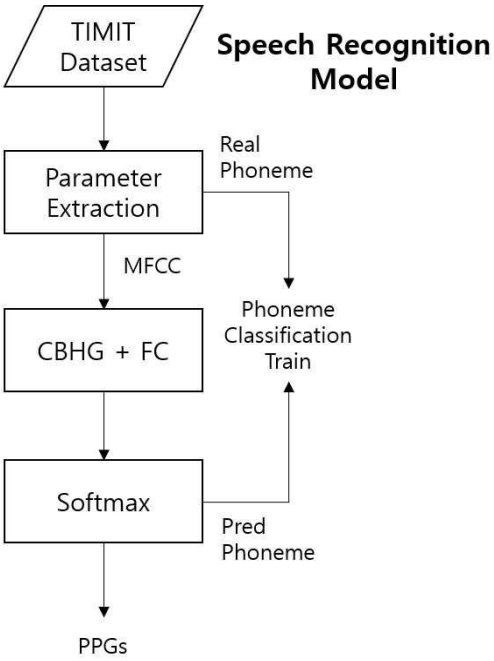
[0083] 본 발명에서는 한 화자에 음성을 다른 화자에 음성 특성에 맞추어 변환하는 음성 변환을 PPGs와 GAN 구조를 이용하여 수행하였는데, 실험결과 Baseline 모델에 비하여 생성 음성에 대해 MOS에서 개선된 결과를 보여주었다.

도면

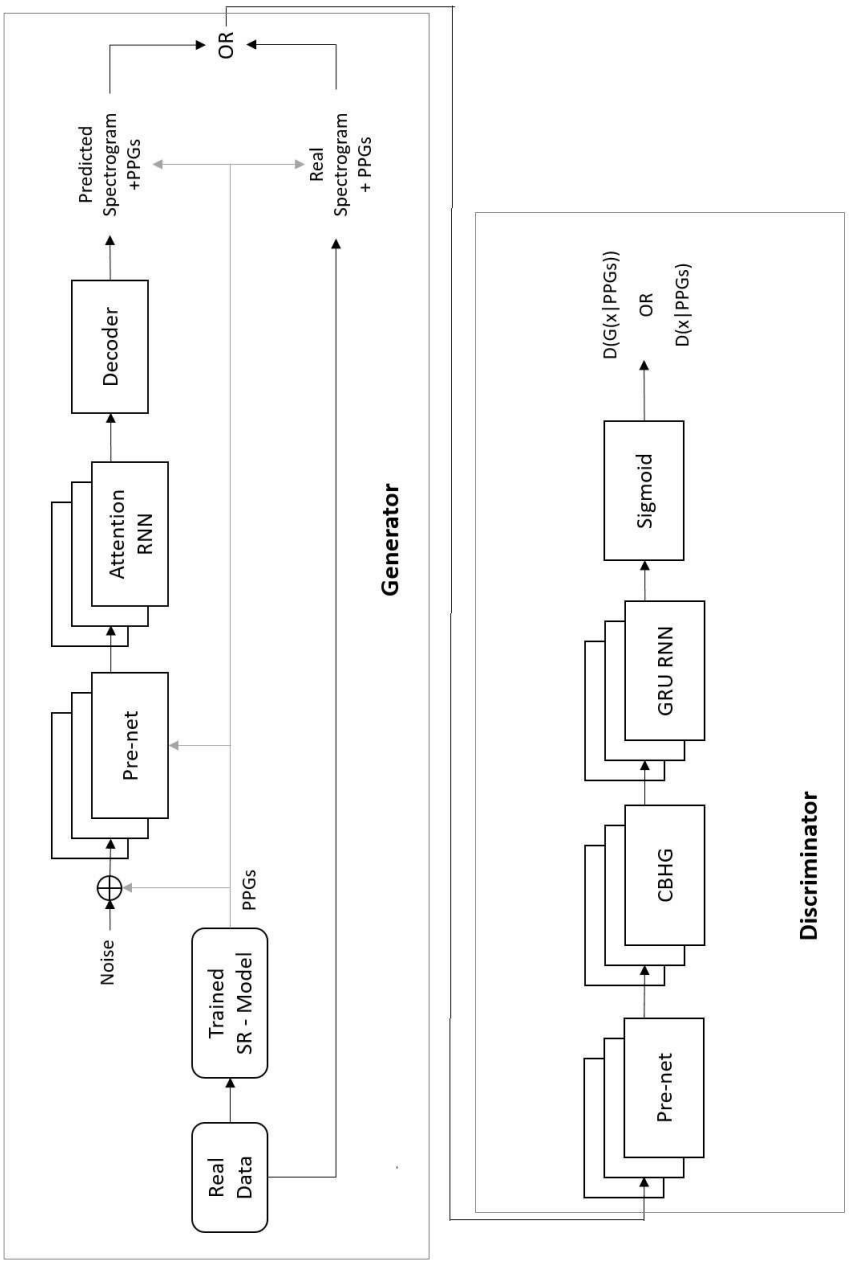
도면1


Speech Recognition Model Training
TIMIT Dataset
Parameter Extraction
MFCC
SR - Model Training
GAN Model Training
Target Speech
Parameter Extraction
MFCC
Trained SR - Model
PPGs
GAN Model Training
Voice Conversion
Source Speech
Parameter Extraction
MFCC
Trained SR - Model
PPGs
Trained Generator
Predicted Spectrogram
Griffin-Lim reconstruction
Converted Speech

도면2


TIMIT Dataset
Speech Recognition Model
Parameter Extraction
Real Phoneme
MFCC
CBHG + FC
Phoneme Classification Train
Softmax
Pred Phoneme
PPGs

도면3


OR
Predicted Spectrogram +PPGs
Real Spectrogram + PPGs
Decoder
Attention RNN
Pre-net
Noise
PPGs
Trained SR - Model
Real Data
Generator
D(G(x|PPGs))
OR
D(x|PPGs)
Sigmoid
GRU RNN
CBHG
Pre-net
Discriminator