논문 19-44-09-02 The Journal of Korean Institute of Communications and Information Sciences '19-09 Vol.44 No.09
https://doi.org/10.7840/kics.2019.44.9.1641

상호 정보량을 이용한 End-to-End 음성 합성에서의 발화 스타일 모델링 기법

이 준 엽◆, 천 성 준*, 최 병 진*, 김 남 수°, 홍 두 화**

Speech Style Modeling Method Using Mutual Information for End-to-End Speech Synthesis

Joun Yeop Lee◆, Sung Jun Cheon*, Byoung Jin Choi*, Nam Soo Kim°, Doo Hwa Hong**

요 약

본 논문에서는 mutual information(MI)를 사용하여 스타일 end-to-end 음성 합성에서 스타일에 텍스트 정보를 없애는 기법을 제안한다. MI을 딥 러닝 환경에서 구현하기 위하여 mutual information neural estimator(MINE)을 활용하였으며 이를 통해 텍스트 정보가 분리된 스타일을 추출하여 음성 합성에 사용할 수 있을 것이다. 제안하는 기법은 VCTK 데이터베이스를 활용하여 실험되었으며 실험 결과 기존의 방식은 Tacotron Global Style Token 기법에 비해 높은 성능을 보임을 확인할 수 있었다.

Key Words : end-to-end speech synthesis, mutual information, mutual information neural estimator, global style token, style modeling

ABSTRACT

In this paper, we propose a novel style modeling method using mutual information(MI) for end-to-end speech synthesis. MI is applied to increase target style information and suppress text information in style embedding by adding MI loss term in objective function. To estimate MI using neural networks, we adopt mutual information neural estimator (MINE). The proposed method was trained using VCTK database and shown to outperform the conventional Tacotron based Global Style Token method in both speech quality and style similarity.

Ⅰ. 서 론

최근 딥러닝을 필두로 한 머신러닝 기법들의 발전과 함께 음성 분야 연구의 성능도 크게 개선되고 있다. 음성 합성 분야 또한 딥러닝 기법의 등장과 함께 비약적인 성능의 발전이 이루어지고 있다. 음성 합성에서의 딥러닝 기법은 기존의 은닉 마르코프 모델 음성 합성[1]과 유사한 형태로 linguistic 특징 벡터를 추출하는 모델, duration 모델, 음향 모델, 보코더 등으로 나누어 근사된 duration을 음향 모델의 입력으로 넣어

※ 이 연구는 방위사업청 및 국방과학연구소의 재원에 의해 설립된 신호정보 특화연구센터 사업의 지원을 받아 수행되었음
◆ First Author : Seoul National University Department of Electrical and Computer Engineering and Institute of New Media and Communications and Human Interface Laboratory, jylee@hi.snu.ac.kr, 학생회원
° Corresponding Author : Seoul National University Department of Electrical and Computer Engineering and Institute of New Media and Communications and Human Interface Laboratory, nkim@snu.ac.kr, 종신회원
* Seoul National University Department of Electrical and Computer Engineering and Institute of New Media and Communications and Human Interface Laboratory, sjcheon@hi.snu.ac.kr, bkchoi@hi.snu.ac.kr, 학생회원
** Independent Personal Researcher, with44u@gmail.com
논문번호 : 201907-127-A-RU, Received July 5, 2019; Revised July 23, 2019; Accepted July 23, 2019

음성 특징 파라미터를 생성하던 방식[2-4]에서, 하나의 딥러닝 구조를 사용하는 end-to-end 음성 합성으로 그 형태가 바뀌어 가며 활발한 연구가 진행되고 있다. 이러한 attention 기반의 end-to-end 음성 합성에는 현재 가장 많이 사용되는 Tacotron[5,6] 기법뿐만 아니라, Deep Voice[7-9], VoiceLoop[10], Char2Wav[11] 등등이 제안되었으며 점점 사람의 음성과 유사한 성능을 보이고 있다. 기존의 End-to-end 음성 합성 모델을 단조로운 낭독체 기반의 단일 화자로 구성된 데이터를 활용하여 학습하는 경우 높은 성능을 보이는 결과가 보고되고 있으나, 스타일 음성 합성으로 표현되는 화자, 혹은 발화 스타일을 바꾸는 음성 합성 기술에 대해서는 아직 많은 연구가 필요한 시점이다.

이러한 스타일 음성 합성은 목표로 하는 스타일의 참조 음성(reference audio)이 있을 때 참조 음성을 인코딩하여 음성 합성 시스템에 이를 반영하도록 하는 방식으로 많은 시도가 일어나고 있다. [12]에서는 운율을 하나의 벡터로 인코딩하는 방식을 사용하였으며 본 논문에서의 베이스라인으로 사용한 global style token(GST)[13] 방식의 경우 참조 음성을 여러 토큰을 사용하여 나타내는 방식을 사용하였다. 이와 같은 기법들은 목표 스타일로 발화하는데 어느 정도 성공을 거뒀지만, 발화 스타일뿐만 아니라 녹음환경을 많이 반영하고, 참조 음성의 길이나 음소 구성과, 합성하고자 하는 문장의 관계에 따라 합성음의 품질이 영향을 많이 받는다는 단점이 있다.

본 논문에서는 이러한 단점을 해결하기 위해 GST Tacotron 시스템을 학습시 원하는 목표 스타일에 집중할 수 있도록 목표 스타일 벡터와 스타일 토큰 층 출력의 상호 정보량을 증가시키는 시스템을 구성하여 이러한 문제점을 해결하고자 한다. 이러한 목표 스타일의 정보량이 증가하는 과정은 목표 스타일 정보량의 증가 측면뿐만 아니라 다른 요소에 의한 영향이 감소되는 효과를 가져올 수 있어 GST Tacotron 의 단점을 보완할 수 있다. 상호 정보량을 계산하기 위해 본 논문에서는 mutual information neural estimation(MINE, 상호 정보량 추정 네트워크) 기법[14]을 사용하여 상호 정보량을 구하고 이를 Tacotron 손실 함수 에 추가하는 방식을 사용하였다. 또한 본 논문의 성능을 검증하기 위하여 크게 화자의 변화를 목표 스타일의 변화로 가정하고 실험을 진행하였다. 실험은 영문을 발음한 108명의 화자로 구성된 VCTK[16] 데이터베이스를 사용하였으며 GST Tacotron 기법과의 비교를 통해 본 논문에서 제안한 방식의 타당성을 검증하였다.

Ⅱ. 스타일 토큰 end-to-end 음성 합성

GST Tacotron [13]에서는 별도의 스타일을 정의하지 않고 학습을 통해 데이터베이스의 스타일들을 딥러닝 구조가 스타일 토큰으로 나타내도록 학습하고, 음성 합성 과정에서 스타일 토큰을 활용하여 참조 음성과 유사한 스타일의 음성을 합성하도록 한다. GST의 전반적인 시스템은 그림 1과 같다. 스타일 토큰을 학습할 시에는 참조 음성이 들어왔을 때 이를 인코더 네트워크를 통과 시킨 후 이 값을 다시 attention과 토큰들로 이루어진 스타일 토큰 층으로 입력해주는 방식으로 스타일을 추출해 낸다. 이렇게 추출한 스타일 임베딩은 Tacotron의 텍스트 인코더 상태에 조건부로 입력되어 Tacotron end-to-end 음성 합성에 적용된다. 스타일은 스타일 토큰들과 각 토큰의 가중치를 통해 조절되며 토큰들은 학습 시 각각 임의로 초기화 된 값을 사용한다. 또한, 학습 시에는 참조 음성과 목표 음성을 동일하게 사용하여 스타일을 학습하며 inference 단계에서는 입력 문장과 상관없는 목표 참조 음성을 사용하게 된다.

Inference 단계는 학습과 동일하게 참조 음성을 사용하는 방법과 토큰의 가중치를 조절하는 방식으로 나눌 수 있다. 스타일 토큰의 경우 학습을 통해 결정되기 때문에 어떤 스타일이 어느 스타일 토큰인지 알기 위해서는 가중치 값을 바꾸어가며 확인하는 작업이 필요하다. 본 논문에서는 전자의 방법을 사용하여 목표 스타일의 음성을 넣어주어 목표 스타일과 유사한 스타일의 음성을 생성하는지 확인해보았다.

GST Tacotron 의 경우 서론에서도 언급하였듯이 목표 스타일이 존재하는 경우에도 이를 반영하는 별도의 방법이 존재하지 않아 학습이 어떤 스타일을 목표로 진행될지 예상하기 어렵다는 문제가 있다.

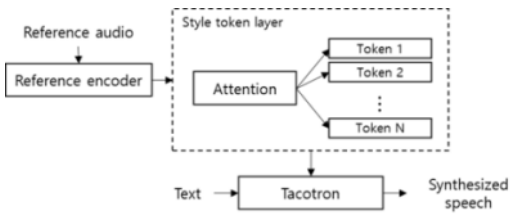

그림 1. Tacotron 시스템 개요도
Fig. 1. Tacotron model diagram

Ⅲ. 상호 정보량 추정 네트워크

여기서는 상호 정보량과 상호 정보량 추정 네트워크에 대해 살펴본다. 본 장에서 살펴본 상호 정보량 추정 네트워크는 다음 장에서 목적 스타일과 스타일 임베딩 사이의 상관관계를 추정하기 위해 사용된다.

3.1 상호 정보량

상호 정보량은 확률 이론과 정보 이론에서 많이 쓰이는 측정값들 중 하나로 엔트로피를 사용하여 2개의 확률 변수들 간의 상호 의존성을 구한다. (X, Y)가 확률변수일 때, 상호 정보량은 다음과 같이 나타낼 수 있다.

$$I(X, Y) := H(X) - H(X|Y) \tag{1}$$

여기서 H는 엔트로피를 의미하며 상호 정보량이 높을수록 X와 Y는 높은 상관관계를 갖는다고 볼 수 있으며 두 확률변수가 독립인 경우에만 상호 정보량 값이 0이 된다. 따라서 상호 정보량은 독립성을 측정하기 위한 척도로 사용되며 이 값이 작을수록 확률변수간의 독립성이 크다는 것을 의미한다. 또한, 상호 정보량은 주변 확률 분포와 결합 분포의 곱을 이용하면 다음 수식과 같이 Kullback-Leibler divergence(KLD)의 형태로 나타낼 수 있다.

$$I(X, Y) := D_{KL}(P_{XY} \| P_X \otimes P_Y) \tag{2}$$

KLD 형태의 상호 정보량과 KLD의 Donsker-Varadhan representation을 사용하면 상호 정보량은 다음과 같은 하계를 가지게 된다.

$$I(X, Y) \geq E_{P_{XY}}[T_\omega] - \log E_{P_X \otimes P_Y}[e^{T_\omega}] \tag{3}$$

여기서 T는 integrability constraint를 만족하는 임의의 함수를 뜻한다.

3.2 상호 정보량 추정 네트워크

상호 정보량은 딥러닝에서의 손실 함수로 사용되기 적절한 척도이지만 연속 확률 변수에서의 계산이 어렵기 때문에 딥러닝에서 널리 사용되지 못하였다. 그러나 MINE은 수식 (3)의 하계를 사용하여 이를 해결하였다. $F = T_{\omega w \in \Omega}$를 심층신경망을 통해 만들어진 통계 네트워크로 불리는 함수들의 집합이라고 할 때, MINE은 다음과 같이 정의할 수 있다.

$$\hat{I}_w(X, Y) = \sup E_{P_{XY}}[T_\omega] - \log E_{P_X \otimes P_Y}[e^{T_\omega}] \tag{4}$$

학습 시 MINE은 결합 분포와 주변 확률 분포로부터 반복적으로 샘플을 뽑고 역전파를 사용하여 통계 네트워크를 업데이트한다. 또한, MINE은 여러 실험을 통해 생성모형을 학습하는데 유용하며 특히, generative adversarial network 등의 학습에 사용되어 높은 성능을 입증하였다[13].

Ⅳ. 상호 정보량을 이용한 end-to-end 음성 합성 시스템

이 장에서는 상호 정보량을 이용한 스타일 end-to-end 음성 합성 시스템을 제안한다. 목표 스타일을 보다 명확히 GST Tacotron 모델에 적용하기 위해서는 먼저 목표하는 스타일에 대한 스타일 벡터가 있어야 한다. 이러한 스타일은 화자, 감정, 발화 속도 등이 될 수 있으며 본 논문에서는 이 중 화자를 목표 스타일로 설정하여 d-vector[15]을 목표 스타일 벡터로 사용하였다.

D-vector는 화자 인식 및 화자 식별 분야에서 가장 널리 사용되는 특징 벡터 중 하나로, 심층신경망을 이용하여 화자를 구별 할 수 있는 특징을 추출하는 기법이다. 학습 단계에서 학습 데이터를 사용하여 프레임 단위의 화자 식별 네트워크를 학습하며 학습된 심층망의 마지막 은닉층의 평균을 취하여 d-vector로 사용한다. D-vector는 화자 정보가 많이 담겨 있어 음성 합성에서 화자 정보를 입력해주기 위하여 시도되어 왔다[17]. 그러나 d-vector는 화자 인식을 위해 사용되는 특징 벡터이므로 음성 합성에 특화되지 못해 성능의 한계를 보였다.

본 논문에서 제안하는 전체적인 시스템 개요는 그림 2와 같다. 목표 스타일을 스타일 토큰에 반영하기 위해서 MINE 네트워크를 통해 스타일 임베딩과 스타일 벡터의 상호 정보량을 계산하여 상호 정보량을 최대화 시킨다. 이를 위해 기존의 GST Tacotron의 손실함수에서 상호정보량을 차감하는 형태를 취함으로써 성능 향상을 시도하였다.

$$Loss = Loss_{Taco} - \alpha I_{MINE}(S, S_v) \tag{5}$$

$Loss_{Taco}$는 기존 GST Tacotron의 손실 함수를

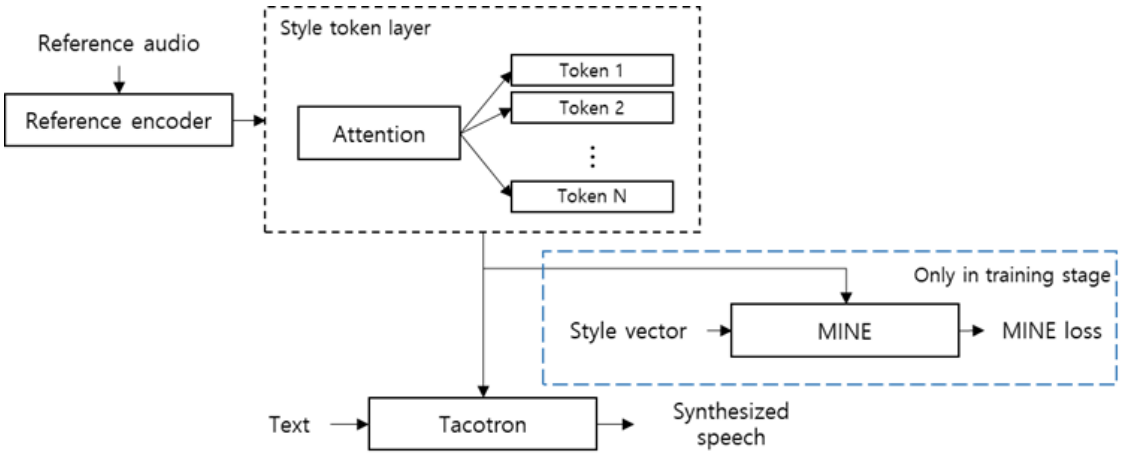

그림 2. 상호 정보량 추정 네트워크를 이용한 시스템 개요도
Fig. 2. Diagram of style end-to-end speech synthesis system using MINE

뜻하며 $I_{MINE}(S, S_v)$는 식 (4)의 추정된 상호 정보량을 뜻하며 S와 S_v는 각각 스타일 임베딩과 스타일 벡터를 뜻한다. 또한, α는 상수 값으로 MINE 손실 값을 얼마나 손실 함수에 반영할지 결정하며 0에서 1의 실수 값을 가지게 설정하여 실험하였다. 상호정보량은 3장에서 보았듯이 독립성을 나타내는 척도이므로 손실 함수를 (5)와 같이 사용하면 목표 스타일 벡터와 GST을 통해 임베딩된 스타일 임베딩 값의 독립성이 줄어들게 되어 목표 스타일 벡터와 스타일 임베딩 값이 통계적으로 의존적이 되도록 학습이 되게 된다. 이는 결과적으로 목표 스타일을 GST Tacotron이 잘 반영되는 것을 의미하여 보다 높은 성능을 기대할 수 있다. 이와 같은 MINE을 통한 손실 함수에서의 상보 정보량 계산을 통하여 본 논문에서는 기존의 GST Tactoron에서 목표하는 스타일이 있어도 이를 효과적으로 반영하지 못하던 점을 효과적으로 보완할 수 있다. 또한, 기존의 d-vector만을 이용할 경우 발생하던 성능의 한계를 스타일 토큰을 이용하여 합성기와 함께 학습함으로써 목표 스타일 정보를 포함하면서 합성에 보다 적합한 스타일 임베딩(화자 임베딩)을 할 수 있다.

또한, 스타일 임베딩에 목표 스타일의 정보량을 극대화 시키는 과정에서 스타일 임베딩에 텍스트의 정보량이 줄어드는 효과가 있기 때문에 스타일 임베딩에 참조 음성의 텍스트에 미치는 영향력을 줄여 보다 안정적으로 참조 음성의 스타일을 반영하는 합성기를 만들 수 있을 것으로 기대된다.

Inference 단계에서는 목표 스타일 벡터를 추가로 사용하여 상호정보량을 계산하지 않으므로 참조 음성만을 사용하며 MINE과 관련된 부분은 생략되어 기존의 GST Tacotron의 방식을 사용하여 간단하게 inference할 수 있다.

Ⅴ. 실험 및 토의

5.1 실험 환경 및 비교 모델

본 논문에서는 VCTK 데이터베이스를 사용하여 두 가지의 비교실험을 진행하였다. 비교 모델은 기존 기법으로는 Ⅱ장의 기본적인 GST Tacotron 구조와 Ⅳ장에서 제안하고 있는 상호 정보량을 이용한 모델이다. 비교 실험은 학습된 합성음의 선호도를 평가한 비교 실험과 목표 스타일과의 유사성의 선호도 평가로 진행되었다. 선호도 평가는 18명의 음성 전문가를 대상으로 실시되었으며 선호되는 음성을 선택하거나 차이가 없는 경우에는 차이 없음(Conv.은 기존 기법을, Prop.는 제안 기법을, 차이 없음은 Same으로 표기)을 선택하도록 하였다. 선호도는 %로 나타내었다. 각 평가에는 40개의 임의의 문장이 사용되었다. 총 네 가지의 스타일(화자)을 평가에 사용하였으며 음질 비교 평가와 유사도 비교 평가의 문장은 다른 문장을 사용하였으며 각 비교 평가 문항은 같은 문장으로 비교하도록 하였다.

VCTK는 영문 낭독체로 이루어져 있으며 총 108명의 화자로 이루어져 있다. 각각의 화자는 다른 환경에서 녹음되었으나 잡음이 없는 환경이며 묵음 구간을 제외하고 약 15분 정도의 음성 길이를 가지고 있다. 또한, 48 kHz의 음성을 22050 Hz로 낮춰 실험을 진행하였고 각 시스템의 음성은 80차의

mel-spectrogram 으로 변환하여 사용하였다. 각 화자에 대한 스타일 벡터인 d-vector의 경우 512차의 d-vector를 추출하여 사용하였다.

학습 문장과 테스트 문장은 다른 문장을 사용하였으며 테스트 시 사용되는 참조 음성 역시 학습에 사용되지 않은 음성을 사용하여 진행하였다. 기본적인 GST Tacotron 구조는 같은 구조를 사용하였으며 다음과 같이 구성되어 있다.

1) Tacotron 구조의 경우 임베딩 순람표를 통해 256차의 텍스트 임베딩을 출력하며 인코더로는 CBHG 모듈을 사용하였다. 또한, feedforward 네트워크를 사용한 prenet과 location sensitive attention을 사용하였다. 디코더는 LSTM 및 LSTM 상태를 mel spectrogram으로 바꿔주는 네트워크를 사용하였다. 또한 후처리로 CBHG를 사용하였으며 최종적으로 Griffin-Lim 보코더를 통해 음성을 합성하였다.

2) 스타일 토큰 층의 attention은 4개의 head를 가진 multi-head attention 구조이며 참조 인코더의 경우 [13]에서의 운율 인코더와 같은 형태의 인코더를 사용하였다.

MINE 네트워크는 2개의 feedforward 은닉층을 가진 네트워크로 구성하였으며 제안한 손실함수에서의 α 값은 사전 실험 결과 성능이 높았던 0.4를 사용하여 실험하였다.

5.2 실험 결과 및 토의

5.2.1 음질 비교 평가

음질 비교 평가의 결과는 그림3과 같다. 각 스타일에 따라 성능의 차이가 있지만 대체로 MINE을 이용하여 학습한 경우 높은 음질 성능을 보임을 확인할 수 있었다. 다만 4번째 스타일의 경우 높은 성능을 보이지 않았는데 그 원인으로는 4번째 스타일 화자의 경우 다른 스타일에 비해 억양이 일반적이지 않으며 참조 음성 안에서의 변화가 다른 스타일에 비해 일관성을 갖지 못해 전반적인 음질이 떨어지기 때문인 것으로 추정 된다. 이와 같이 스타일 유사도뿐만 아니라 음질 측면에서의 향상이 존재하는 원인은 기존 기법에서는 학습 시에 스타일 임베딩과 합성기 출력 음성에 동일한 문장을 사용하기 때문에 스타일 임베딩에 텍스트에 의한 영향이 클 수 있다. 그러나 본 논문에서 제안한 기법을 이용한다면 목표 스타일에 대한 정보량을 극대화 시키는 과정에서 텍스트에 의한 영향이 줄기 때문인 것으로 추정된다.

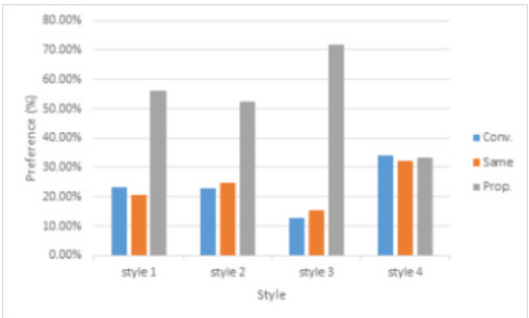

그림 3. 음질 성능 비교 평가 그래프
Fig. 3. Subjective test result graph

5.2.2 스타일 유사도 비교 평가

스타일 유사도 비교 평가의 결과는 그림 4와 같다. 스타일 유사도의 비교 평가 결과 기존의 방식에 비해 상호 정보량을 사용하여 학습한 경우 모든 스타일에 대해서 더 높은 성능을 보였다. 특히 음질 비교 평가에서 결과가 안 좋았던 스타일 4의 경우 오히려 스타일 유사도에서는 가장 높은 성능 차이를 보였다. 스타일 4의 음색이 일반적이지 않은 것을 감안하였을 때 기존의 방식으로는 이러한 목표 스타일이 잘 모델링 되지 않았지만 제안 방식으로는 목표 스타일이 잘 모델링 된 것으로 분석할 수 있을 것이다.

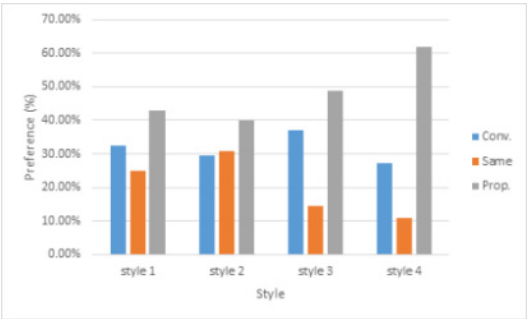

그림 4. 스타일 유사도 비교 평가 그래프
Fig. 4. Subjective test(style similarity) result graph

Ⅵ. 결 론

본 논문에서는 상호정보량을 이용하여 GST Tacotron을 학습하는 방법을 제안하고 실험을 통해 검증하였다. 목표로 하는 특정 스타일에 대하여 상호 정보량을 증대시키는 방향으로 학습시킴으로써 목표 스타일을 더 잘 반영할 수 있는 스타일 음성 합성기를 만들 수 있었다. 또한 스타일의 유사도 측면뿐만 아니라 음질 자체에 대한 성능 향상 또한 확인할 수 있었다.

추후 연구에서는 텍스트의 영향을 상호 정보 추정 네트워크를 통해 스타일에서 제거함으로써 제안 방식에 비해 텍스트에 영향을 받지 않는 스타일 음성 합성을 만들 것이다.

References

[1] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system version 2.0," in *Proc. ISCA SSW6*, Bonn, Germany, Aug. 2007.

[2] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process*, pp. 7962-7966, Vancouver, Canada, May 2013.

[3] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process*, pp. 4470-4474, Brisbane, Australia, Apr. 2015.

[4] Z. Wu and S. King, "Investigating gated recurrent networks for speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process*, pp. 5140-5144, Shaghai, China, Mar. 2016.

[5] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, pp. 4006-4010, Stockholm, Sweden, Aug. 2017.

[6] J. Shen, R. Pang, R. J Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process*, pp. 4779-4783, Calgary, Canada, Apr. 2018.

[7] S. O. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, J. Raiman, S. Sengupta, and M. Shoeybi, "Deep voice: Real-time neural text-to-speech," *CoRR*, vol. abs/1702.07825, 2017.

[8] S. O. Arik, G. F. Diamos, A. Gibiansky, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," *CoRR*, vol. abs/1705.08947, 2017.

[9] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: 2000-speaker neural text-to-speech," *CoRR*, vol. abs/1710.07654, 2017.

[10] Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani, "VoiceLoop: Voice fitting and synthesis via a phonological loop," *ICLR*, Vancouver, Canada, Apr. 2018.

[11] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, "Char2Wav: End-to-end speech synthesis," *ICLR*, Toulon, France, Apr. 2017.

[12] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron," *ICML*, Stockholm, Sweden, Jul. 2018.

[13] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," *ICML*, Stockholm, Sweden, Jul. 2018.

[14] I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, A. Courville, and R. D. Hjelm, "Mine: mutual information neural estimation," *ICML*, Stockholm, Sweden, Jul. 2018.

[15] L. Wan, et al., "Generalized end-to-end loss for speaker verification," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process*, pp. 4879-4883, Calgary, Canada, Apr. 2018.

[16] *CSTR VCTK Corpus*, retrieved, Jun. 2019, https://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html

[17] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno, and Y. Wu, “Transfer learning from speaker verification to multispeaker text-to-speech synthesis,” in *Proc. Advances in Neural Inf. Process. Syst.*, pp. 4485-4495, Montreal, Canada, Dec. 2018.

이 준 엽 (Joun Yeop Lee)

2013년 2월 : 서울대학교 전기정보공학부 학사 졸업
2013년 3월~현재 : 서울대학교 전기컴퓨터공학부 석박사통합 과정 박사과정
<관심분야> 음성 신호 처리, 음성 합성, 뉴럴 네트워크
[ORCID:0000-0002-3316-4808]

천 성 준 (Sung Jun Cheon)

2013년 8월 : 서울대학교 전기정보공학부 학사 졸업
2014년 3월~현재 : 서울대학교 전기정보공학부 석박사통합 과정 박사과정
<관심분야> 음성 신호 처리, 음성 합성, 뉴럴 네트워크
[ORCID:0000-0002-7293-6997]

최 병 진 (Byoung Jin Choi)

2013년 5월 : University of Wisconsin - Madison 전기공학부 학사 졸업
2017년 3월~현재 : 서울대학교 전기정보공학부 석박사통합 과정 박사과정
<관심분야> 음성 신호 처리, 음성 합성, 뉴럴 네트워크
[ORCID:0000-0003-1319-8215]

김 남 수 (Nam Soo Kim)

1988년 : 서울대학교 전자공학과 학사 졸업
1990년 : 한국과학기술원 전기 및 전자공학과 석사 졸업
1994년 : 한국과학기술원 전기 및 전자공학과 박사 졸업
1993년~1998년 : 삼성종합기술원 전문연구원
1998년~현재 : 서울대학교 전기정보공학부 교수
<관심분야> 음성 신호처리, 음성 인식, 통계적 신호처리, 패턴 인식, 휴먼 인터페이스
[ORCID:0000-0002-0568-4902]

홍 두 화 (Doo Hwa Hong)

2007년 8월 : KAIST 전기및전자공학전공 학사 졸업
2015년 2월 : 서울대학교 전기컴퓨터공학 박사 졸업
<관심분야> 음성 합성, 데이터 사이언스, 딥러닝
[ORCID:0000-0002-5503-4828]