pISSN 2005-8063
eISSN 2586-5854
2024. 3. 31.
Vol.16 No.1
pp. 67-76

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2024.16.1.067

Check for updates

One-shot multi-speaker text-to-speech using RawNet3 speaker representation*

Sohee Han · Jisub Um · Hoirin Kim**

Department of Electrical and Electronic Engineering,
Korea Advanced Institute of Science and Technology, Daejeon, Korea

Abstract

Recent advances in text-to-speech (TTS) technology have significantly improved the quality of synthesized speech, reaching a level where it can closely imitate natural human speech. Especially, TTS models offering various voice characteristics and personalized speech, are widely utilized in fields such as artificial intelligence (AI) tutors, advertising, and video dubbing. Accordingly, in this paper, we propose a one-shot multi-speaker TTS system that can ensure acoustic diversity and synthesize personalized voice by generating speech using unseen target speakers' utterances. The proposed model integrates a speaker encoder into a TTS model consisting of the FastSpeech2 acoustic model and the HiFi-GAN vocoder. The speaker encoder, based on the pre-trained RawNet3, extracts speaker-specific voice features. Furthermore, the proposed approach not only includes an English one-shot multi-speaker TTS but also introduces a Korean one-shot multi-speaker TTS. We evaluate naturalness and speaker similarity of the generated speech using objective and subjective metrics. In the subjective evaluation, the proposed Korean one-shot multi-speaker TTS obtained naturalness mean opinion score (NMOS) of 3.36 and similarity MOS (SMOS) of 3.16. The objective evaluation of the proposed English and Korean one-shot multi-speaker TTS showed a prediction MOS (P-MOS) of 2.54 and 3.74, respectively. These results indicate that the performance of our proposed model is improved over the baseline models in terms of both naturalness and speaker similarity.

Keywords: speech synthesis, multi-speaker text-to-speech (TTS), speaker embedding, speaker adaptation, one-shot speech synthesis

1. 서론

음성합성(text-to-speech, TTS) 시스템은 입력 텍스트가 주어졌을 때, 이에 대한 음성을 합성해주는 시스템이다. 음성합성 시스템은 크게 2가지 과정으로 이뤄진다. 첫 번째는 음향 모델로 텍스트를 입력으로 받아 mel-spectrogram이나 MFCC(mel-frequency cepstral coefficient) 같은 중간 단계의 feature를 생성하는 과정이다. 두 번째로는 보코더 모델을 통해서 중간 단계의 feature들을

* This work was supported by the National Research of Foundation of Korea (No. 2021R1A2C1014044).
** hoirkim@kaist.ac.kr, Corresponding author
Received 31 January 2024; Revised 21 March 2024; Accepted 21 March 2024

raw waveform 형태로 변환한다. 본 연구에서도 음향 모델과 보코더로 이뤄진 음성합성 시스템을 구현하였다.

최근 음성합성 기술의 발전으로 인해 생성된 합성음의 음질과 자연성이 크게 향상되었고, 사람의 목소리처럼 더 자연스러운 음성을 합성할 수 있는 수준에 이르렀다. 이러한 발전 덕에 음성합성 기술은 음성 자동 응답 서비스(automatic response service, ARS), AI(artificial intelligence) 영어 튜터, 콘텐츠 더빙, 언어장애 치료 등 다양한 분야에 활용되고 있다. 그렇기에 다양한 목소리 선택지를 제공할 수 있는 음성합성 모델과 개인의 선호에 맞춘 커스텀 보이스 생성 기술에 대한 필요성은 지속적으로 증가하고 있다.

이러한 필요성에 따라 개인 맞춤형 음성합성 모델과 관련된 연구들이 최근 들어 많은 관심을 받고 있다. 이와 관련된 연구들로는 few-shot multi-speaker TTS, speaker conditional TTS, speaker adaptation TTS 등이 있다(Cooper et al., 2020; Moss et al., 2020; Zhao et al., 2022). Few-shot multi-speaker TTS 모델은 적은 데이터셋만을 활용함에도 원하는 화자의 목소리 특성이 반영된 음성을 생성할 수 있는 다화자 음성합성 시스템이다.

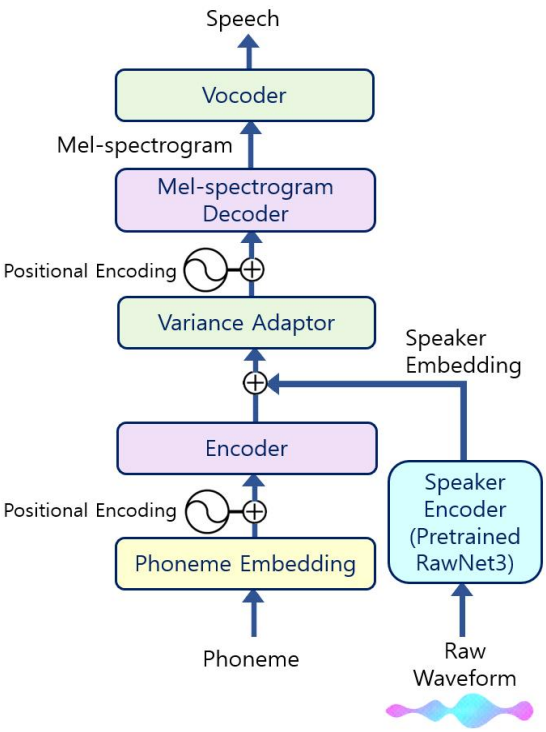
Hsu et al.(2018)의 GMVAE 모델은 VAE(variational autoencoder) framework와 GMM(Gaussian mixture model) 기반의 latent distribution을 사용하여 음질이 향상되고, 발화 스타일의 조절이 가능한 TTS 모델을 제안했다. Cooper et al.(2020)의 연구는 LDE(learnable dictionary encoding) 기반의 화자 임베딩과 angular softmax 손실함수로 few-shot multi-speaker TTS 모델을 구현하였다. Choi et al.(2020)의 Attentron 모델은 적은 데이터만으로 화자 목소리 특성을 반영하기 위해 2가지 화자 인코더를 제안했다. 각각의 인코더로는 fine grained 인코더와 coarse grained 인코더를 사용했다. Fine grained 인코더는 여러 목표 음성 파일을 잘 활용하고, 일반화 능력을 개선한, 가변 길이의 임베딩을 다루는 모듈이다. Coarse grained 인코더는 목표 음성의 전반적인 특성을 포함한 global 임베딩을 생성하는 모듈로, 동일 화자 음성들의 감정이나 운율로 인한 가변적 요소를 안정화하는 역할을 한다. Casanova et al.(2021)의 SC-GlowTTS는 GE2E loss기반 SV(speaker verification) 모델의 LSTM(long short-term memory) 층과 선형 층의 수를 늘린 구조를 화자 인코더로 활용했다(Wan et al., 2018). 또한 화자 인코더를 25,000명 화자로 구성된 데이터로, Angular Prototypical loss로 훈련해 화자 유사도를 개선했다. Casanova et al.(2022)의 YourTTS는 Voxceleb2 데이터셋과 손실함수 prototypical angular loss와 softmax loss로 사전훈련된 ResNet34 기반 화자인식 모델을 화자 인코더로 활용하여 당시 speaker adaptation TTS 연구에서 최고 성능을 기록했다(Heo et al., 2020). 또한 4차원 언어 임베딩을 통해 4개 국어로 확장했다.

본 연구에서는 이전 연구들에 비해 개선된 화자 인코더 모델을 사용함으로써 화자 유사도를 개선한 '원샷 다화자 음성합성 모델'을 제시한다. 구체적으로 Jung et al.(2022)의 화자 식별 모델인 RawNet3를 통해 추출한 화자 표현을 활용하여, 훈련 데이터셋에 포함되지 않은 화자의 목소리로도 음성을 합성할 수 있는 다화자 음성합성 모델을 구현하였다. 본 연구의 가장 주된 목표는 음향적 다양성을 확보함으로써 활용 분야를 넓히는 것과 훈련에 사용한 데이터셋의 화자 수보다 훨씬 많은 화자 목소리의 구현이 가능하도록 함으로써 훈련의 효율성을 증대하는 것이다. 또한 본 연구에서는 영어뿐만 아니라 한국어에 대해서도 원샷 다화자 음성합성 시스템을 구현하였다. 비교 모델로는 제안 모델과 같은 음향 모델에 GE2E(generalized end-to-end) loss 기반 화자 식별 모델과 ResNet34 기반 화자 인식 모델을 각각 결합한 모델들을 활용했다. 이러한 비교 모델들과의 주관적 평가와 객관적 평가에서 제안 모델이 화자 유사도와 자연성 측면에서 향상된 성능을 보였다(Kwon et al., 2021; Wan et al., 2018).

2. One-Shot Multi-Speaker Text-to-Speech 제안 모델

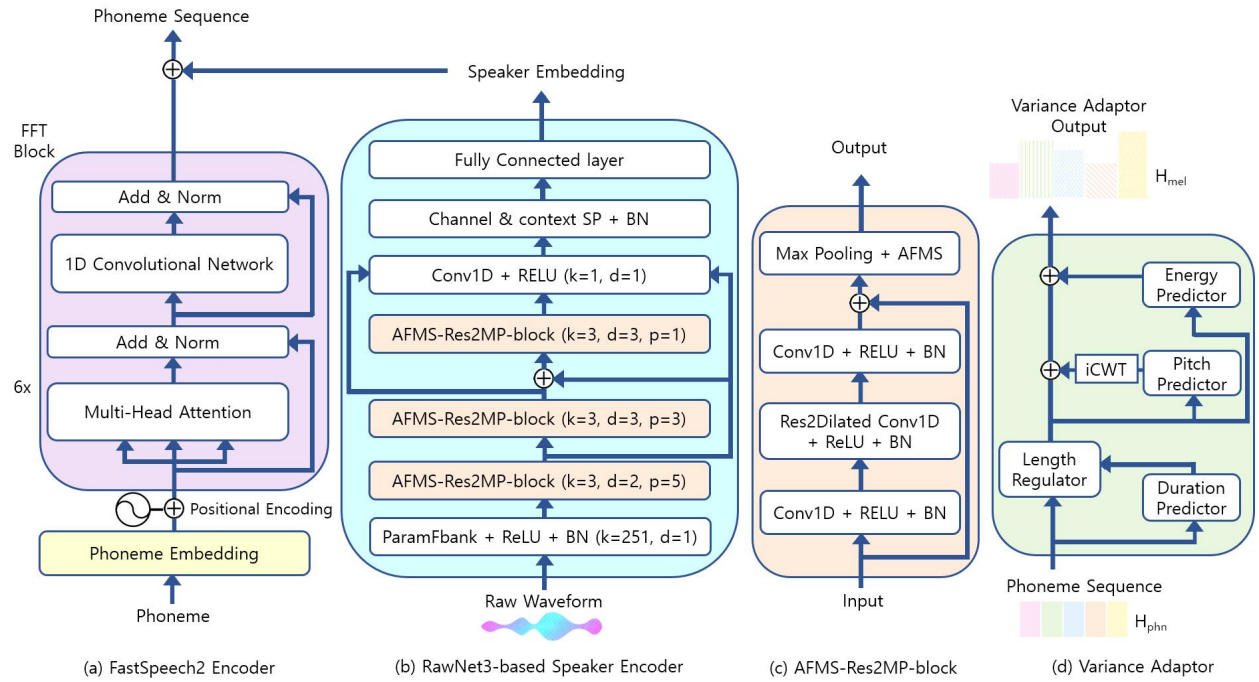
2.1. 모델 구조

본 연구에서는 'RawNet3를 통해 추출한 화자 표현을 활용한 원샷 다화자 음성합성' 모델을 제안한다(그림 1). 원샷 다화자 음성합성 모델을 구현하기 위해 기존에 연구되었던 다화자 음성합성 모델에 화자 인코더 모듈을 결합하였다. TTS 모델로 FastSpeech2 음향 모델과 HiFi-GAN 보코더를 사용하였고, 사전훈련된 RawNet3를 화자 인코더로 결합하여 구현하였다(Jung et al., 2022; Kong et al., 2020; Ren et al., 2020).



TTS, text-to-speech.

그림 1. 제안하는 원샷 다화자 음성합성 모델
Figure 1. Proposed one-shot multi-speaker TTS model



AFMS, extended feature map scaling; Res2MP, Res2Net with max pooling; iCWT, inverse continuous wavelet transform; ParamFbank, parameterized filterbanks.

그림 2. 음향 모델의 인코더, 화자 인코더, variance adaptor의 구조
(a) FastSpeech2의 인코더, (b) RawNet3 기반의 화자 인코더, (c) RawNet3의 AFMS-Res2MP Block , (d) FastSpeech2의 variance adaptor
Figure 2. Architecture of acoustic model encoder, speaker encoder and variance adaptor
(a) Encoder of FastSpeech2, (b) RawNet3-based speaker encoder, (c) AFMS-Res2MP block of RawNet3, (d) Variance adaptor of FastSpeech2

2.1.1. 음향 모델

음향 모델은 입력으로 텍스트를 받아 mel-spectrogram을 생성하는 역할을 한다. 본 연구에서는 인코더, 디코더, variance adaptor로 구성된 FastSpeech2를 기본 모델로 활용한다(Ren et al., 2020). 인코더는 phoneme embedding sequence를 입력으로 받아 phoneme hidden sequence로 변환하는 역할을 한다[그림 2(a)]. 인코더는 6개의 FFT(feed-forward transformer) 블록으로 구성되고, 각각의 FFT 블록은 multi-head attention과 2개의 1D-컨볼루션 층과 ReLU 활성화 함수로 구성된다. Multi-head attention은 여러 개의 head로 cross-position에 대한 정보를 얻는 데 활용된다.

FastSpeech2의 FFT 블록은 FastSpeech의 FFT 블록과 동일한 구조를 사용하였다(Li et al., 2019; Ren et al., 2019). FastSpeech의 FFT 블록은 transformer TTS의 인코더 구조를 기반으로 하였지만, multi-head attention과 FFN(feed-forward network)로 이뤄진 Transformer TTS 인코더 블록에서 FFN 대신 2개의 1D-컨볼루션 층과 ReLU로 대체한 구조를 가진다. Transformer TTS에서 FFN 내부의 fully connected 층을 시퀀스 데이터 처리에 강점을 가진 1D-컨볼루션 층으로 대체함으로써 문맥 정보를 더 잘 파악하도록 하였다. 이는 가까이 위치한 hidden state들이 더 높은 관련성을 갖도록 한 것이기에 기존 Transformer TTS의 인코더 구조보다 음성합성 모델 구현에 더 적합한 구조이다.

인코더 출력인 phoneme hidden sequence에 화자 임베딩이 더해져 phoneme sequence가 생성되고, 이는 인코더 다음에 이어지는 variance adaptor의 입력으로 들어간다. '그림 2(d)'의 variance adaptor는 입력 sequence에 duration, pitch, energy에 해당하는 variance 정보를 반영해준다. 훈련 과정에서는 ground-truth인 duration, pitch, energy 값을 활용하고 추론 과정에서는 해당 값들을 사용할 수 없기에 각각의 정보를 예측해주는 predictor module로 예측한 값을 활용한다. 해당 predictor module은 세 개의 하위 predictor로 구성되고, 각각의 구조는 모두 같다. 2개의 1D-컨볼루션 층과 ReLU 활성화 함수와 선형 층으로 구성되고, 3개의 층 사이에 layer normalization과 dropout layer가 존재한다.

Variance adaptor의 출력으로 얻게 되는 variance 정보가 반영된 phoneme sequence는 디코더의 입력으로 들어간다. 디코더는 해당 입력을 바탕으로 mel-spectrogram을 생성하는 역할을 한다. 구조는 인코더와 마찬가지로 6개의 FFT 블록으로 구성된다. 디코더를 통해 생성된 mel-spectrogram은 HiFi-GAN 보코더의 입력으로 사용되고, 보코더는 출력으로 raw waveform 형태의 음성을 생성한다(Kong et al., 2020).

2.1.2. 화자 인코더

Few-shot 다화자 음성합성 모델은 적은 수의 목표 화자의 발화에서 화자 임베딩을 추출하고, 해당 목소리 특성을 반영해 음성을 생성하는 시스템이다. 이를 구현하기 위해 음향 모델에 화자 임베딩을 추출할 수 있는 화자 인코더를 결합한 구조를 활용한다. 화자 인코더는 목표 음성에서 억양, 악센트, 운율 등의 전반적인 화자 음색 정보를 얻어내는 역할을 한다. '서론'에서 살펴봤듯이, few-shot TTS 모델에서 화자 인코더로는 주로 화자

인식(speaker recognition, SR), 화자 식별(speaker verification, SV) 시스템이 활용된다(Casanova et al., 2021, 2022).

본 연구에서는 화자 인코더로, raw waveform을 입력으로 하는 모델 중 우수한 화자 식별 성능을 가진 RawNet3를 화자 인코더로 활용해, 유사한 화자 음색 표현에 강점을 가진 '원샷 다화자 음성합성' 모델을 구현하였다(Jung et al., 2022). 그림 2(b)의 RawNet3는 RawNet2와 ECAPA-TDNN 기반의 구조로 이루어진다. RawNet2 모델의 filterbank 구조를 사용하되, 실수 기반에서 복소수 기반으로 확장하였다(Desplanques et al., 2020; Jung et al., 2020). ECAPA-TDNN의 SE-Res2 블록과 유사한 구조를 가진, RawNet3의 AFMS-Res2MP 블록에서는 RawNet2의 AFMS (extended feature map scaling)를 적용한다는 차이가 있다[그림 2(c)]. 여기서 SE-Res2는 SE(squeeze-excitation)와 Res2Net로 이뤄진 구조를, AFMS-Res2MP는 AFMS, Res2Net, max pooling이 결합된 구조를 지칭한다. FMS(feature map scaling)는 여러 필터마다 독립적인 scale vector를 부여하여 CNN(convolutional neural network)의 feature map에서 discriminative representation을 추출하는 방법이다. 화자를 식별하는 데 있어 SE보다 FMS를 적용하는 것이 높은 정확도를 보인다(Hu et al., 2018).

RawNet3의 입력으로 들어온 raw waveform은 pre-emphasis와 instance normalization 단계를 거친 후 parameterized analytic filterbanks를 통해 time-frequency representation으로 변환된다. Filterbanks의 출력은 세 개의 AFMS-Res2MP 블록들의 입력이고, 세 개의 블록들에서 첫 번째 블록과 두 번째 블록의 출력은 세 번째 블록의 출력에 더해진다. AFMS-Res2MP 블록은 1D-컨볼루션 층과 ReLU, Res2Dilated 1D-컨볼루션 층과 ReLU, 1D-컨볼루션 층과 ReLU, AFMS 층으로 구성된다. AFMS-Res2MP 블록의 출력은 1D-컨볼루션 층과 ReLU 활성화 함수, 채널 및 문맥 종속적 통계 풀링 층을 통과한다. 최종적으로 fully-connected 층을 거치면 256차원 화자 임베딩이 출력으로 나온다. 이러한 화자 임베딩은 그림 1과 같이 TTS 모델에 결합하여 활용한다.

2.1.3. 보코더

보코더는 mel-spectorgram을 waveform으로 변환해주는 시스템이다. 본 연구에서는 좋은 음질의 음성을 생성하고 합성 속도가 빠른 HiFi-GAN 모델을 보코더로 활용한다(Kong et al., 2020). HiFi-GAN 모델은 합성음의 음질과 합성 효율성 사이의 trade-off와 관련해 세 개의 generator 버전이 존재하고, V1 generator 모델을 중심으로 경량화한 generator 모델들을 V2, V3로 지칭한다. 본 연구에서는 합성음의 음질에 중점을 둔 generator V1을 활용했고, generator V1은 hidden unit size가 512이고, kernel unit size가 [16, 16, 4, 4], kernel radius가 [3, 7, 11] 이다.

HiFi-GAN은 GAN(generative adversarial network) 구조 기반의 모델로 generator와 discriminator로 구성된다. Generator는 입력으로 들어온 mel-spectrogram을 raw waveform으로 변환시켜주는 역할을 한다. 해당 모델은 다양한 kernel size와 dilation size를 가지는 MRF(multi-receptive field fusion) 모듈이 활용되어 다양한 길이의 패턴 정보를 얻을 수 있다.

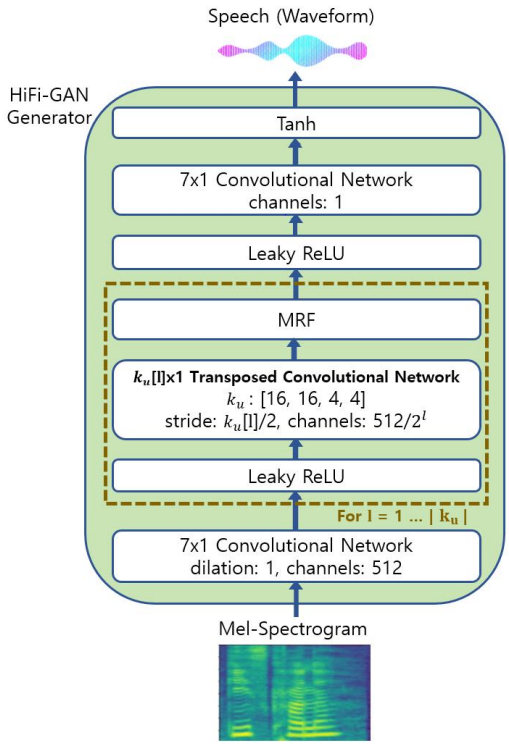


그림 3. HiFi-GAN 보코더의 generator
Figure 3. Generator of HiFi-GAN vocoder

Discriminator는 MPD(multi-period discriminator)와 MSD(multi-scale discriminator)로 구성된다. MPD는 5개의 하위 discriminator로 구성되고, 각각은 입력 오디오의 특정 주기에 따라 샘플들을 처리하는 역할을 한다. 이때 주기는 [2, 3, 5, 7, 11]로 설정하여 overlap이 되지 않도록 하였다. MSD는 세 개의 하위 discriminator로 구성되고, 각각의 discriminator는 2와 4의 factor로 down-sampling된 음성을 입력으로 받는다. 이를 통해 음성을 다양한 스케일로 처리해 각각 다른 주파수에 대해 학습할 수 있게 된다.

Generator와 discriminator는 GAN loss, mel-spectrogram loss, feature matching loss로 구성된 final loss를 통해 훈련되고, generator와 discriminator의 final loss 각각의 수식은 다음과 같다.

$$L_G = \sum_{k=1}^{K} [L_{Adv}(G; D_k) + \lambda_{fm} L_{FM}(G; D_k)] + \lambda_{mel} L_{Mel}(G) \quad (1)$$

$$L_D = \sum_{k=1}^{K} L_{Adv}(D_k; G) \quad (2)$$

2.2. 훈련

훈련에 사용할 데이터의 전처리 단계에서는 duration, pitch, energy에 대한 ground truth variance 값을 추출하는 과정을 거친다. 음성이 지속되는 시간을 나타내는 duration은 MFA(Montreal forced alignment)를 통해 ground truth 값을 구한다. 감정과 운율 정보를 담고 있는 pitch는 world vocoder를 활용해 pitch contour를 추출하고, CWT(continuous wavelet transform)로 변환하여 ground truth 값을 얻는다(Morise et al., 2016). Mel-spectrogram의

frame-level 진폭으로서 발화 음량을 나타내는 energy는 STFT (short time Fourier transform)를 통해 ground truth 값을 구한다.

제안 모델인 원샷 다화자 음성합성 모델을 구현하는 과정에서 화자 인코더로는 사전 훈련된 RawNet3 모델을, 보코더로는 사전 훈련된 HiFi-GAN을 사용하였다. 사전 훈련된 RawNet3는 Voxceleb1, 2 데이터셋으로, 사전 훈련된 HiFi-GAN 보코더는 LJSpeech, VCTK, LibriTTS 데이터셋으로 훈련된 모델이다 (Chung et al., 2018; Nagrani et al., 2017). 훈련 과정에서는 훈련 데이터셋의 발화에 존재하는 화자 표현을 얻는 데 RawNet3 기반의 화자 인코더가 활용된다. 구체적으로 raw waveform에서 화자 특징이 담긴 화자 임베딩을 추출한다.

FastSpeech2 음향 모델의 훈련은 데이터 전처리 단계에서 구한 phoneme duration, pitch, energy, mel-spectrogram에 대한 ground truth 값들을 활용해서 진행한다. 훈련은 (1) 인코더와 디코더로 구성된 전체 음향 모델 훈련, (2) variance adaptor 내부의 predictor 들의 훈련으로, 크게 두 과정으로 나눌 수 있다.

우선 인코더와 디코더로 구성된 전체 음향 모델을 훈련하는 과정에서는 텍스트와 발화로 구성된 다화자 음성 데이터셋이 활용된다. 음향 모델의 입력은 텍스트이고, 가장 먼저 인코더를 거쳐 hidden phoneme sequence가 나온다. 이 sequence에 RawNet3 기반의 화자 인코더로 추출한 화자 임베딩을 더해 phoneme sequence를 구한다. 이 phoneme sequence는 variance adaptor의 입력으로 들어간다. 훈련 과정에서 variance adaptor는 입력 sequence에 phoneme duration, pitch, energy 각각에 대한 ground truth 값이 더해진 출력을 만든다. 이후 variance adaptor의 출력은 디코더로 들어가 mel-spectrogram으로 변환된다. 최종적으로, 디코더를 통해 구한 mel-spectrogram과 raw waveform에 STFT를 적용해 구한 ground truth mel-spectrogram 사이의 MAE(mean absolute error) loss로 정의되는 L_{recon}을 통해 학습된다[식 (3)].

$$L_{recon} = \frac{1}{n}\sum_{i=1}^{n} | Y_i - \hat{Y}_i | \quad (3)$$

Variance adaptor를 구성하고 있는 세 개의 predictor들은 모두 전처리 과정에서 구한 ground truth 값과 각각의 predictor들로 예측한 값 사이의 MSE(mean squared error) loss로 정의되는 $L_{variance}$를 통해 훈련이 진행된다[식 (4)]. Ground truth 값은 각 frame에 대해 256차원으로 quantize된 임베딩 형태이다.

$$L_{variance} = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \quad (4)$$

전체 원샷 다화자 음성합성 모델을 훈련하는 L_{total}은 음향 모델을 훈련하는 과정의 손실함수인 L_{recon}과 variance adaptor 내부 predictor의 훈련에 활용되는 $L_{variance}$의 합으로 정의되며, 모델 전체는 jointly한 방식으로 최적화된다[식 (5)].

$$L_{total} = L_{recon} + L_{variance} \quad (5)$$

2.3. 추론

추론 과정에서는 앞선 단계에서 훈련된 원샷 다화자 음성합성 시스템을 활용하여, 입력 텍스트에 대응되는 음성을 생성한다. 훈련된 음향 모델에 텍스트와 목표 화자의 발화가 입력으로 들어온다. 입력 텍스트는 음향 모델의 인코더를 통과해 phoneme hidden sequence로 만들어진다. 해당 phoneme hidden sequence에, 목표 화자의 발화에서 RawNet3 기반 화자 인코더를 사용해 원샷 방식으로 추출한 화자 임베딩을 더함으로써 화자 특성을 반영한다. 그 후 variance adaptor를 통과하는데, 훈련 과정에서와 달리 추론 과정에는 variance adaptor 내부의 predictor들로 예측한 값들을 활용한다. 이를 통해 음의 높낮이, 속도, 진폭 정보를 반영한다. 이후 variance adaptor의 출력은 디코더 입력으로 들어가 mel-spectrogram이 출력으로 나온다. 디코더의 출력은 사전 훈련된 HiFi-GAN 보코더를 거쳐 raw waveform으로 변환되고, 최종적으로 목표 화자의 음색이 반영된 음성이 생성된다.

3. 실험

3.1. 실험 환경

본 연구에서는 영어뿐만 아니라 한국어에 대해서도 원샷 다화자 음성합성 시스템을 구현하였다. 영어 다화자 데이터셋으로는 LibriTTS-360과 Libri-100을 사용했다(Zen et al., 2019). 영어 데이터셋인 LibriTTS는 Project Gutenberg의 텍스트 파일과 LibriVox에서 자원봉사자들이 녹음한 음성 파일들로 구성된 데이터셋이다. 영어 원샷 다화자 음성합성 모델의 훈련에 사용한 Libri-360은 남녀 904명의 화자가 포함된 총 360시간 분량, 추론에 사용한 Libri-100은 남녀 247명의 화자가 포함된 총 100시간 분량의 데이터셋이다. LibriTTS-360과 Libri-100은 중복되지 않는 데이터셋이다. 한국어 다화자 데이터셋으로는 남녀 31명의 화자가 속한 총 30시간 분량의, 전문 성우들이 녹음한 음성 파일들로 구성된 자체 수집 데이터셋을 활용했다. 이 한국어 다화자 데이터셋은 Korean multi-speaker DB로 표기한다. 한국어 원샷 다화자 음성합성 모델의 훈련에는 Korean multi-speaker DB의 31중 27명 화자의 데이터셋을, 추론에는 훈련과 겹치지 않는 4명 화자의 데이터셋을 활용하였다. 두 언어의 데이터셋 모두 22.05kHz로 다운 샘플링하여 활용했다. 80차원의 mel-spectrogram을 추출할 때 FFT size는 1,024, hop size는 256, window size는 1,024로 설정했다. 훈련하는 과정의 초기 learning rate는 $2e^{-4}$ (0.0002)로 설정하였으며, Vaswani et al.(2017)에서와 같은 learning rate 스케줄러를 사용하였다.

3.2. 모델 설정

3.2.1. Baseline 모델

본 논문에서는 제안한 원샷 다화자 음성합성 모델의 성능을

표 1. 실험에 사용한 음향 모델의 하이퍼파라미터
Table 1. Hyperparameters of acoustic model

Hyperparameter	Value
Phoneme embedding dimension	256
Encoder layers	4
Encoder hidden	256
Encoder Conv1D kernel	9
Encoder Conv1D filter size	1,024
Encoder attention heads	2
Mel-spectrogram decoder layers	4
Mel-spectrogram decoder hidden	256
Mel-spectrogram decoder Conv1D kernel	9
Mel-spectrogram decoder Conv1D filter size	1,024
Mel-spectrogram decoder attention headers	2
Encoder / decoder dropout	0.1
Variance predictor Conv1D kernel	3
Variance predictor Conv1D filter size	256
Variance predictor dropout	0.5
Waveform decoder convolution blocks	30
Waveform decoder dilated Conv1D kernel size	3
Waveform decoder transposed Conv1D filter size	64
Waveform decoder skip channel Size	64
Batch size	32

평가하기 위해 세 개의 비교 모델을 설정하였다. 세 개의 시스템은 FastSpeech2와 Speaker ID, FastSpeech2와 GE2E loss 기반 화자 식별 모델, FastSpeech2와 ResNet34기반 화자 인식 모델이고, 이는 각각 FastSpeech2+Speaker ID, FastSpeech2+GE2ESV, FastSpeech2+ResNet34SE로 표기하였다. 음향 모델로는 모두 같은 FastSpeech2 모델과 하이퍼파라미터 값을 사용하였고, 보코더로는 사전 훈련된 HiFi-GAN 보코더를 동일하게 활용하였다.

Speaker ID는 화자 임베딩 테이블을 기반으로 하고 있으며, 정수 형태의 입력을 넣어서 대응되는 화자 임베딩에 접근하는 방식이다. GE2E loss 기반의 화자 식별 모델인 GE2ESV는 LSTM으로 구성된 TE2E(text-to-end-to-end) SV 모델을 GE2E loss 기반으로 훈련함으로써 generalization 능력을 개선한 모델이다(Kwon et al., 2021). ResNet34 기반 화자 인식 모델인 ResNet34SE는 잡음에 강인한 성능을 가진 시스템이고, ResNet34와 TDNN 모델 기반의 구조이다. 이 모델은 컨볼루션 층과 4개의 residual 블록, flatten 층, angular prototypical 손실함수와 vanilla softmax 손실함수를 결합한 ASP 층, 선형 층으로 구성된다(Kwon et al., 2021).

비교 모델들의 화자 인코더는 모두 Voxceleb1, 2 데이터셋으로 사전 훈련된 모델이다. 이러한 화자 인코더들의 출력은 모두 화자 임베딩이고, 각각 화자 임베딩을 음향 모델 인코더의 출력인, 256차원의 hidden phoneme sequence에 맞추어 더해준다.

3.2.2. 제안 모델

제안하는 원샷 다화자 음성합성 모델은 FastSpeech2 음향 모델, RawNet3 기반 화자 인코더, HiFi-GAN 보코더로 구성된다. FastSpeech2 음향 모델은 phoneme embedding layer, 6개의 FFT 블록으로 이뤄진 인코더, variance adaptor, 6개의 FFT 블록으로 이뤄진 디코더로 구성된다. 인코더와 디코더 구조에서의 FFT 블록은 self-attention 층과 1D 컨볼루션 층으로 구성되어 있다. Variance adaptor 내부의 predictor는 모두 같은 구조이고, 1D-컨볼루션 2층과 ReLU, layer normalization, dropout 층, 선형 층으로 구성되어 있다. 실험에 사용한 음향 모델의 하이퍼파라미터 설정은 표 1에 제시하였다.

화자 인코더인 RawNet3는 parameterized analytic filterbanks와 세 개의 AFMS-Res2MP 블록, 1D-컨볼루션 층과 ReLU 활성화 함수, 채널 및 문맥 종속적 통계 풀링 층, fully-connected 층으로 구성된다. RawNet3의 출력은 256차원의 화자 임베딩이다.

3.3. 평가 지표

평가 지표 본 논문에서는 제안 모델의 성능을 평가하기 위하여 주관적 평가 지표와 객관적 평가 지표를 활용하였다. 주관적 평가 지표로 MOS(mean opinion score)를 사용하였고, 그중에서도 NMOS(naturalness mean opinion score), SMOS(similarity mean opinion score)를 사용하였다. 객관적 평가 지표로는 P-MOS(prediction mean opinion score), SECS(speaker embedding cosine similarity)를 사용하였다. 또한 t-SNE(t-distributed stochastic neighbor embedding)를 활용하여, 원음, 비교 모델, 제안 모델을 통해 생성한 합성음 각각에서 추출한 화자 임베딩을 2차원 공간에 시각적으로 제시하였다.

4. 실험 결과

4.1. 주관적 평가

주관적 평가는 한국어 원샷 다화자 음성합성 모델에 대해서만 평가를 진행하였다. 평가 지표로는 1~5점 사이의 0.5점 단위로, 9-scale로 평가하는 MOS를 사용하였다. 본 논문의 제안 모델에 대해 자연성과 화자 유사도 각각을 평가하기 위해 NMOS와 SMOS를 사용했다. NMOS는 합성음의 자연스러운 정도를 평가하는 지표이고, SMOS는 목표 화자의 음성과 합성음을 비교하는 방식으로 화자 유사도를 평가하는 지표다.

평가자로는 20대에서 50대까지의 정상 청력을 가진 남녀 21명이 참여하였다. 훈련 중 봤던 화자, 훈련 중 보지 않았던 화자에 대한 한국어 원샷 다화자 음성합성 결과, 크게 두 범주로 나눠 각각 합성음 40문장에 대해 NMOS와 SMOS 평가를 진행하였다. 이와 함께 신뢰 구간 95%에 대한 값도 산출하여 제시하였다.

표 2는 한국어 원샷 다화자 음성합성 모델의 주관적 평가 결과이다. 훈련 중 본 화자에 대한 MOS 결과에서 제안 모델의 합성음에 대한 NMOS는 비교 모델과 달리 3.5점을 넘은 3.66점이고, SMOS는 3.97점으로 4점에 근접한 높은 점수를 보였다.

훈련 중 보지 않은 화자에 대한 합성음 평가 결과는 훈련 중 본 화자에 대한 합성음보다 전반적으로 낮은 점수 분포를 보이지만, 제안 모델이 자연성과 화자 유사도 측면에서 가장 높은 점수를 보인다는 점에서 같은 경향성을 나타낸다. 이를 종합하면, NMOS를 통해 제안 모델의 합성음이 가장 자연스러움을 알 수 있고, SMOS를 통해 화자 음색을 가장 잘 반영하고 있음을 알 수 있다.

표 2. 한국어 원샷 다화자 음성합성 모델의 주관적 평가 결과
Table 2. Subjective evaluation results of Korean one-shot multi-speaker TTS

[TTS+화자 인코더] 모델	Seen		Unseen	
	NMOS	SMOS	NMOS	SMOS
Ground truth	4.34±0.06	4.72±0.02	4.32±0.04	4.72±0.02
FastSpeech2+Speaker ID	3.33±0.08	3.58±0.07	ND	ND
FastSpeech2+GE2ESV	3.48±0.06	3.67±0.07	3.12±0.05	2.83±0.07
FastSpeech2+ResNet34SE	3.28±0.08	3.43±0.13	3.23±0.05	3.03±0.07
(Proposed) FastSeech2+RawNet3	3.66±0.06	3.97±0.04	3.36±0.04	3.16±0.04

TTS, text-to-speech; NMOS, naturalness mean opinion score; SMOS, similarity MOS; GE2E, generalized end-to-end; ND, not detected.

표 3. 원샷 다화자 음성합성 모델의 객관적 평가 결과
Table 3. Objective evaluation results of one-shot multi-speaker TTS

[TTS+화자 인코더] 모델	English				Korean			
	Seen		Unseen		Seen		Unseen	
	P-MOS	SECS	P-MOS	SECS	P-MOS	SECS	P-MOS	SECS
Ground truth	3.53±0.06	0.98	3.80±0.02	0.98	4.09±0.04	0.99	3.85±0.03	0.99
FastSpeech2+Speaker ID	2.97±0.05	0.85	ND	ND	3.45±0.05	0.93	ND	ND
FastSpeech2+GE2ESV	2.94±0.05	0.88	2.44±0.02	0.89	3.57±0.05	0.94	3.43±0.03	0.85
FastSpeech2+ResNet34SE	2.75±0.05	0.78	2.29±0.02	0.80	3.52±0.05	0.93	3.35±0.02	0.83
(Proposed) FastSeech2+RawNet3	3.00±0.05	0.90	2.54±0.03	0.89	3.58±0.05	0.94	3.74±0.02	0.87

TTS, text-to-speech, P-MOS, prediction MOS; SECS, speaker embedding cosine similarity; GE2E, generalized end-to-end; ND, not detected.

4.2. 객관적 평가

P-MOS는 딥러닝 모델을 활용하여 합성음의 자연성 및 전반적인 음질에 대해 객관적으로 평가하는 방법이다. P-MOS는 주관적 평가인 MOS가 가진, 평가자들 성향이나 평가 시점에 따라 결과가 달라질 수 있는 한계점을 극복한 객관적 평가 방법이다.

SECS는 합성음이 목표 음성과 얼마나 화자 음색이 유사한지를 평가하는 방법이다. SECS 값은 합성음과 원음 각각에서 화자 임베딩을 추출하여 코사인 유사도를 계산한 값으로 –1에서 1 사이의 값을 가지고, 1에 가까울수록 화자 유사도가 높은 화자 임베딩임을 의미한다.

표 3의 영어 다화자 음성합성 모델로 생성한 합성음의 P-MOS 평가 결과를 통해, 훈련 중 본 화자와 훈련 중 보지 않은 화자 두 측면에 대해서 모두, 제안 모델이 가장 음질이 우수하고, 자연스러운 합성음을 출력함을 알 수 있다. 특히 제안 모델로 훈련 중 본 화자에 대해 합성한 음성은 다른 모델들과 달리 ground truth와 같은 3점대로 높은 P-MOS 점수를 보였다. 또한 제안 모델의, 훈련 중 본 화자와 보지 않은 화자 각각에 대한 합성음의 SECS 평가 결과가 0.9, 0.89임을 통해, 제안 모델이 목표 화자의 음색 정보들을 더 잘 반영하고 있다는 것을 확인할 수 있다.

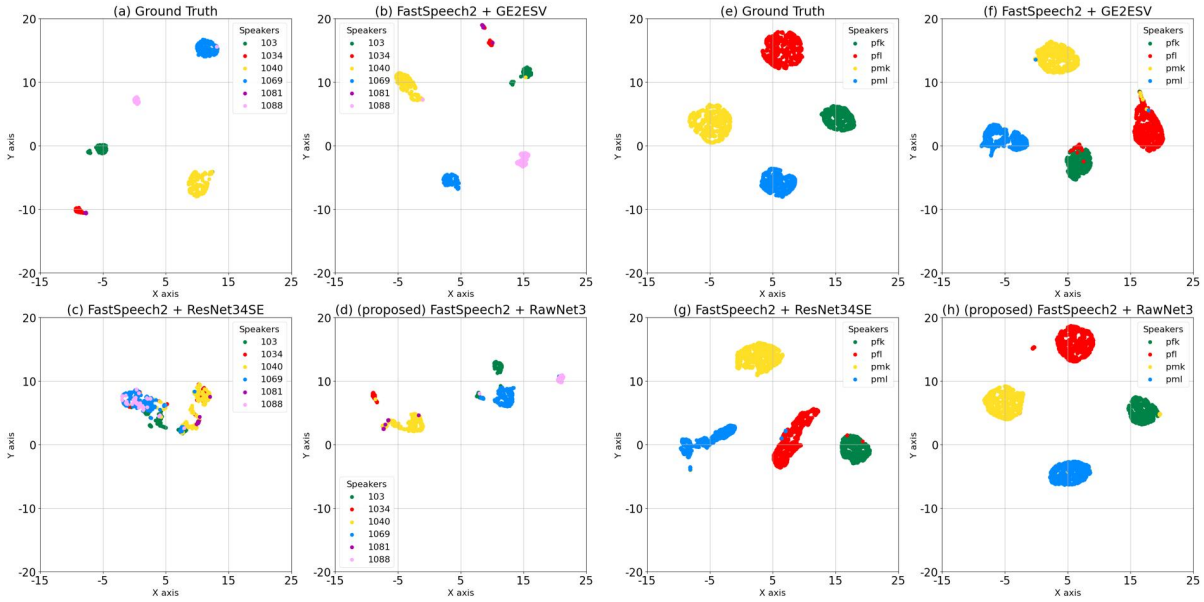
표 3의 한국어 다화자 음성합성 모델로 생성한 합성음의 P-MOS 평가 결과를 통해, 훈련 중 본 화자와 훈련 중 보지 않은 화자에 대한 합성음 모두, 제안 모델로 합성한 음성이 자연성 및 전반적인 음질이 가장 우수함을 알 수 있다. 특히 훈련 중 보지 않은 화자에 대한 P-MOS는 원음과 비교했을 때도 0.1점의 근소한 차이를 보인다. 또한 훈련 중 본 화자와 보지 않은 화자, 각각에 대한 제안 모델의 합성음의 SECS가 비교군보다 높은 것을 통해, 화자 음색을 표현하는 데 강점이 있음을 알 수 있다.

4.3. 시각화

t-SNE는 높은 차원의 데이터를 저차원으로 투영하면서 데이터 간의 관계를 시각적으로 나타내는 도구이다. 본 논문에서는 원음과 원샷 음성합성 모델들의 합성음 각각에서 추출한 화자 임베딩을 고차원에서 저차원으로 대응시켜, 화자 임베딩에 대한 정보를 시각적으로 파악할 수 있도록 하였다. 합성음에 대한 t-SNE plot에서 다른 화자에 대한 임베딩을 잘 구분할수록, 원음의 화자 임베딩 plot과 비슷한 양상일수록 화자 특징이 적절하게 추출되어 합성 모델에 반영되었음을 의미한다. 영어 원샷 다화자 음성합성 모델의 plot은 그림 3의 (a)~(d)에 나타내었고, 한국어 다화자 음성합성 모델의 plot은 그림 3의 (e)~(h)에 나타내었다. 그림 3의 (a)~(d)와 (e)~(h) 각각에 (1) ground truth(원음), (2) FastSpeech2+GE2ESV(비교 모델), (3) FastSpeech2+ResNet34SE(비교 모델), (4) 제안 모델을 순서대로 제시하였다.

그림 3의 (a)~(d)는 영어 원샷 다화자 음성합성 모델을 통해 생성한, 훈련 중 보지 않은 화자에 대한 합성음의 화자 임베딩 t-SNE plot이다. 해당 plot에는 LibriTTS-100 중 6명 화자에 대한 데이터셋의 텍스트를 기반으로 만든 합성음이 활용되었다.

그림 3에서 한국어 원샷 다화자 음성합성 모델에 대한 t-SNE plot인 (e)~(h)에는 Korean multi-speaker DB 중 4명의 화자에 대한 데이터셋의 텍스트를 입력하여 생성한 합성음이 사용되었다. 해당 한국어 시스템에 대한 plot을 통해, 제안 모델은 비교 모델들보다 분류하는 데 있어 오류율이 낮고, 같은 화자에 대해 밀집된 형태의 군집을 이루는 화자 임베딩이 음성 합성하는 데 활용되었음을 확인할 수 있다. 이를 통해, 제안 모델로 동일한 화자에 대한 발화를 합성하는 경우, 다른 비교 모델들보다 더 안정적인 성능을 기대할 수 있다.



t-SNE, t-distributed stochastic neighbor embedding; TTS, text-to-speech.

그림 4. 원샷 다화자 음성합성 모델의 t-SNE. (a)~(d) 영어, (e)~(h) 한국어
Figure 4. t-SNE of one-shot multi-speaker TTS. (a)~(d) English, (e)~(h) Korean

5. 결론

본 논문에서는 훈련 중 보지 않은 화자에 대해서 원샷 방식으로 음성합성이 가능한 '원샷 다화자 음성합성 시스템'을 제안하였다. 해당 시스템은 RawNet3 기반 화자 인코더와 음성합성 모델을 결합한 구조를 활용해 구현하였다. 이를 통해 훈련 중 보지 않은 화자를 포함해 여러 화자에 대한 음성합성이 가능하게 되어, 음향적 다양성을 확보할 수 있었다. 또한 훈련 데이터보다 화자 음색 표현 가능 범위를 확장하여 모델의 효율성을 높였다.

성능 평가는 영어와 한국어 원샷 다화자 음성합성 모델에 대해 진행하였다. 결과를 살펴보면, 합성음의 자연성을 평가하는 주관적 평가 지표 NMOS와 객관적 평가 지표 P-MOS에서 모두, 제안 모델의 합성음이 비교 모델들의 합성음보다 높은 점수를 보였다. 또한 합성음의 화자 유사성을 평가하는 주관적 평가 지표인 SMOS와 객관적 평가 지표 SECS에서도, 제안 모델의 합성음이 가장 우수한 화자 유사도를 나타냈다. 이를 통해 제안 모델이 비교군에 비해 자연성과 화자 유사성 모두 향상된 모델임을 알 수 있었다. 또한 시각화 도구인 t-SNE로는 제안 모델로 생성한 합성음의 화자 임베딩 plot이 원음의 화자 임베딩 plot과 유사한 형태를 보임을 확인하였다.

위의 실험 결과들을 통해, 본 논문에서 제안한 모델인 'RawNet3로부터 추출한 화자 임베딩을 활용한 one-shot multi-speaker TTS 시스템'이 다른 화자 인코더 기반의 비교 모델들보다, 훈련 과정에서 본 적이 없는 화자에 대해서도 더 나아진 화자 유사도와 자연성을 가진 음성 생성이 가능함을 입증하였다. 또한 본 연구 결과는 다양한 목소리 선택지를 제공하거나 원하는 커스텀 보이스를 생성하는 기술로 활용될 수 있기에 의미가 있는 연구이다.

References

Cooper, E., Lai, C. I., Yasuda, Y., Fang, F., Wang, X., Chen, N., & Yamagishi, J. (2020, May). Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings. *Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6184-6188). Barcelona, Spain.

Choi, S., Han, S., Kim, D., & Ha, S. (2020, October). Attentron: Few-shot text-to-speech utilizing attention-based variable-length embedding. *Proceedings of the Interspeech 2020.* Shanghai, China.

Casanova, E., Shulby, C., Gölge, E., Müller, N. M., de Oliveira, F. S., Candido A. C. Jr., Soares, A. S., ... & Ponti, M. A. (2021, August-September). Sc-glowTTS: An efficient zero-shot multi-speaker text-to-speech model. *Proceedings of the Interspeech 2021* (pp. 3645-3649). Brno, Czechia.

Casanova, E., Weber, J., Shulby, C., Candido Jr., A., Gölge, E., & Ponti, M. A. (2022, June). Yourtts: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. *Proceedings of the 39th International Conference on Machine Learning* (pp. 2709-2720). Baltimore, MD.

Chung, J. S., Nagrani, A., & Zisserman, A. (2018, September). VoxCeleb2: Deep speaker recognition. *Proceedings of the Interspeech* (pp. 1086-1090). Hyderabad, India.

Desplanques, B., Thienpondt, J., & Demuynck, K. (2020, October). ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. *Proceedings of*

the Interspeech (pp. 3830-3834). Shanghai, China.

Heo, H. S., Lee, B. J., Huh, J., & Chung, J. S. (2020, October). Clova baseline system for the VoxCeleb speaker recognition challenge 2020. *Proceedings of the Interspeech*. Shanghai, China.

Hu, J., Shen, L., & Sun, G. (2018, June). Squeeze-and-excitation networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7132-7141). Salt Lake City, UT.

Hsu, W. N., Zhang, Y., Weiss, R. J., Zen, H., Wu, Y., Wang, Y., Cao, Y., ... Pang, R. (2018, April－May). Hierarchical generative modeling for controllable speech synthesis. *Proceedings of the International Conference on Learning Representations*. Vancouver, BC.

Jung, J., Kim, Y., Heo, H. S., Lee, B. J., Kwon, Y., & Chung, J. S. (2022, September). Pushing the limits of raw waveform speaker recognition. *Proceedings of the Interspeech 2022* (pp. 2228-2232). Incheon, Korea,

Jung, J., Kim, S., Shim, H., Kim, J., & Yu, H. (2020, October). Improved rawnet with feature map scaling for text-independent speaker verification using raw waveforms. *Proceedings of the Interspeech 2020* (pp. 1496-1500). Shanghai, China.

Kwon, Y., Heo, H. S., Lee, B. J., & Chung, J. S. (2021, June). The ins and outs of speaker recognition: Lessons from VoxSRC 2020. *Proceedings of the ICASSP 2021－2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5809-5813). Toronto, ON.

Kong, J., Kim, J., & Bae, J. (2020). HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, *33*, 17022-17033.

Li, N., Liu, S., Liu, Y., Zhao, S., & Liu, M. (2019, July). Neural speech synthesis with transformer network. *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 6706-6713). Honolulu, HI.

Morise, M., Yokomori, F., & Ozawa, K. (2016). WORLD: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Transactions on Information and Systems*, *99*(7), 1877-1884.

Moss, H. B., Aggarwal, V., Prateek, N., González, J., & Barra-Chicote, R. (2020, May). Boffin TTS: Few-shot speaker adaptation by Bayesian optimization. *Proceedings of the ICASSP 2020－2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7639-7643). Barcelona, Spain.

Nagrani, A., Chung, J. S., & Zisserman, A. (2017, August). VoxCeleb: A large-scale speaker identification dataset. *Proceedings of the Interspeech 2017* (pp. 2616-2620). Stockholm, Sweden.

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T. Y. (2019). FastSpeech: Fast, robust and controllable text to speech. *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*. Vancouver, BC.

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T. Y. (2020, June). FastSpeech 2: Fast and high-quality end-to-end text to speech. Retrieved from arXiv:2006.04558v8

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., ... Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. Long Beach, CA.

Wan, L., Wang, Q., Papir, A., & Moreno, I. L. (2018, April). Generalized end-to-end loss for speaker verification. *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4879-4883). Calgary, AB.

Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., Chen, Z., ... Wu, Y. (2019, September). LibriTTS: A corpus derived from librispeech for text-to-speech. *Proceedings of the Interspeech 2019*. Graz, Austria.

Zhao, B., Zhang, X., Wang, J., Cheng, N., & Xiao, J. (2022, May). nnspeech: Speaker-guided conditional variational autoencoder for zero-shot multi-speaker text-to-speech. *Proceedings of the ICASSP 2022－2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4293-4297). Singapore.

• **한소희 (Sohee Han)**
한국과학기술원 전기및전자공학부
대전광역시 유성구 대학로 291
Tel: 042-350-7617
Email: hansh@kaist.ac.kr, 123445679@naver.com
관심분야: 음성합성

• **엄지섭 (Jisub Um)**
한국과학기술원 전기및전자공학부
대전광역시 유성구 대학로 291
Tel: 042-350-7617
Email: twiz0311@kaist.ac.kr
관심분야: 음성합성, 음색 변환

• **김희린 (Hoirin Kim)** 교신저자
한국과학기술원 전기및전자공학부
대전광역시 유성구 대학로 291
Tel: 042-350-7417
Fax: 042-350-7619
Email: hoirkim@kaist.ac.kr
관심분야: 음성인식, 음성합성, 화자인식, 패턴인식

RawNet3를 통해 추출한 화자 특성 기반 원샷 다화자 음성합성 시스템*

한 소 희 · 엄 지 섭 · 김 회 린

한국과학기술원 전기 및 전자공학부

국문초록

최근 음성합성(text-to-speech, TTS) 기술의 발전은 합성음의 음질을 크게 향상하였으며, 사람의 음성에 가까운 합성음을 생성할 수 있는 수준에 이르렀다. 특히, 다양한 음성 특성과 개인화된 음성을 제공하는 TTS 모델은 AI(artificial intelligence) 튜터, 광고, 비디오 더빙과 같은 분야에서 널리 활용되고 있다. 따라서 본 논문은 훈련 중 보지 않은 화자의 발화를 사용하여 음성을 합성함으로써 음향적 다양성을 보장하고 개인화된 음성을 제공하는 원샷 다화자 음성합성 시스템을 제안했다. 이 제안 모델은 FastSpeech2 음향 모델과 HiFi-GAN 보코더로 구성된 TTS 모델에 RawNet3 기반 화자 인코더를 결합한 구조이다. 화자 인코더는 목표 음성에서 화자의 음색이 담긴 임베딩을 추출하는 역할을 한다. 본 논문에서는 영어 원샷 다화자 음성합성 모델뿐만 아니라 한국어 원샷 다화자 음성합성 모델도 구현하였다. 제안한 모델로 합성한 음성의 자연성과 화자 유사도를 평가하기 위해 객관적인 평가 지표와 주관적인 평가 지표를 사용하였다. 주관적 평가에서, 제안한 한국어 원샷 다화자 음성합성 모델의 NMOS(naturalness mean opinion score)는 3.36점이고 SMOS(similarity MOS)는 3.16점이었다. 객관적 평가에서, 제안한 영어 원샷 다화자 음성합성 모델과 한국어 원샷 다화자 음성합성 모델의 P-MOS(prediction MOS)는 각각 2.54점과 3.74점이었다. 이러한 결과는 제안 모델이 화자 유사도와 자연성 두 측면 모두에서 비교 모델들보다 성능이 향상되었음을 의미한다.

핵심어: 음성합성, 다화자 음성합성, 화자 임베딩, 화자 적응, 원샷 음성합성

* 본 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2021R1A2C1014044).