The Journal of the Convergence on Culture Technology (JCCT)
Vol. 7, No. 4, pp.675-681, November 30, 2021. pISSN 2384-0358, eISSN 2384-0366

http://dx.doi.org/10.17703/JCCT.2021.7.4.675
JCCT 2021-11-83

x-vector를 이용한 다화자 음성합성 시스템

A Multi-speaker Speech Synthesis System Using X-vector

조민수*, 권철홍**

Min Su Jo*, Chul Hong Kwon**

요 약 최근 인공지능 스피커 시장이 성장하면서 사용자와 자연스러운 대화가 가능한 음성합성 기술에 대한 수요가 증가하고 있다. 따라서 다양한 음색의 목소리를 생성할 수 있는 다화자 음성합성 시스템이 필요하다. 자연스러운 음성을 합성하기 위해서는 대용량의 고품질 음성 DB로 학습하는 것이 요구된다. 그러나 많은 화자가 발화한 고품질의 대용량 음성 DB를 수집하는 것은 녹음 시간과 비용 측면에서 매우 어려운 일이다. 따라서 각 화자별로는 소량의 학습 데이터이지만 매우 많은 화자의 음성 DB를 사용하여 음성합성 시스템을 학습하고, 이로부터 다화자의 음색과 운율 등을 자연스럽게 표현하는 기술이 필요하다. 본 논문에서는 화자인식 기술에서 사용하는 딥러닝 기반 x-vector 기법을 적용하여 화자 인코더를 구성하고, 화자 인코더를 통해 소량의 데이터로 새로운 화자의 음색을 합성하는 기술을 제안한다. 다화자 음성합성 시스템에서 텍스트 입력에서 멜-스펙트로그램을 합성하는 모듈은 Tacotron2로, 합성음을 생성하는 보코더는 로지스틱 혼합 분포가 적용된 WaveNet으로 구성되어 있다. 학습된 화자 임베딩 신경망에서 추출한 x-vector를 Tacotron2에 입력으로 추가하여 원하는 화자의 음색을 표현한다.

주요어 : x-vector, 화자 임베딩, 다화자, 음성합성

Abstract With the recent growth of the AI speaker market, the demand for speech synthesis technology that enables natural conversation with users is increasing. Therefore, there is a need for a multi-speaker speech synthesis system that can generate voices of various tones. In order to synthesize natural speech, it is required to train with a large-capacity. high-quality speech DB. However, it is very difficult in terms of recording time and cost to collect a high-quality, large-capacity speech database uttered by many speakers. Therefore, it is necessary to train the speech synthesis system using the speech DB of a very large number of speakers with a small amount of training data for each speaker, and a technique for naturally expressing the tone and rhyme of multiple speakers is required. In this paper, we propose a technology for constructing a speaker encoder by applying the deep learning-based x-vector technique used in speaker recognition technology, and synthesizing a new speaker's tone with a small amount of data through the speaker encoder. In the multi-speaker speech synthesis system, the module for synthesizing mel-spectrogram from input text is composed of Tacotron2, and the vocoder generating synthesized speech consists of WaveNet with mixture of logistic distributions applied. The x-vector extracted from the trained speaker embedding neural networks is added to Tacotron2 as an input to express the desired speaker's tone.

Key words : X-vector, Speaker embedding, Multi-speaker, Speech synthesis

*정회원, ㈜사운드잇 연구원 (제1저자)
**정회원, 대전대학교 정보통신·전자공학과 교수 (교신저자)
접수일: 2021년 9월 6일, 수정완료일: 2021년 9월 19일
게재확정일: 2021년 9월 28일

Received: September 6, 2021 / Revised: September 19, 2021
Accepted: September 28, 2021
*Corresponding Author: chkwon@dju.ac.kr
Dept. of Information, Communication, Electronics Engineering, Daejeon Univ, Korea

Ⅰ. 서 론

음성합성(Text-to-Speech, TTS) 시스템은 입력으로 들어온 텍스트로부터 스펙트로그램을 생성하고, 이를 이용하여 음성신호를 합성하여 출력으로 내보낸다. 기존의 음성합성 시스템은 텍스트 정규화, 구문 분석, G2P(Grapheme-to-Phoneme) 변환, 운율 분석, 발성 길이 및 피치 조절, 음향 특징 생성 및 합성음 생성 등 여러 개의 모듈로 구성되어 있다[1]. 이에 반해, 딥러닝 기반 종단형(End-to-End) 음성합성 방식은 전체 과정을 하나 또는 둘로 통합하여 처리하는데, 일반적으로 텍스트에서 멜-스펙트로그램을 합성하는 과정과 멜-스펙트로그램에서 합성음을 생성하는 보코더로 구성되어 있다[2].

딥러닝 기술을 적용한 음성합성 시스템은, 1인 화자 음성합성 시스템인 경우 합성음의 품질이 사람이 녹음한 음성 품질과 유사할 정도로 발전하고 있다. 따라서 현재 상용화된 대부분의 음성합성 시스템은 대용량의 음성 데이터를 사용하여 딥러닝 모델을 학습하고 이로부터 고품질의 합성음을 생성하는 방식을 적용한다. 그러나 이를 위해서는 15시간 이상의 매우 많은 학습용 음성 DB가 필요하여, 음성 DB 녹음 및 정제 시간과 비용 문제로 소수 화자의 음색만을 표현할 수 있다.

최근 인공지능 스피커 시장이 성장하면서 사용자와 자연스러운 대화를 할 수 있는 음성합성 기술에 대한 수요가 증가하고 있다. 따라서 1인 화자의 합성음을 생성하는 것을 넘어 다양한 음색의 목소리를 생성할 수 있는 다화자 음성합성 기술이 필요하다.

자연스러운 음성을 합성하기 위해서는 대용량의 고품질 음성 DB로 학습하는 것이 요구되며, 다화자 음성합성을 위해서는 화자마다 매우 많은 양의 음성 데이터가 필요하다. 많은 화자가 발화한 고품질의 대용량 음성 DB를 수집하는 것은 녹음 시간과 비용 측면에서 매우 어려운 일이고, 이를 대량으로 확보해야 되는 상황은 연구개발의 진입장벽을 크게 높이는 요인이 된다. 따라서 각 화자별로는 소량의 학습 데이터이지만 매우 많은 화자의 음성 DB를 사용하여 음성합성 시스템을 학습하고, 이로부터 다화자의 음색과 운율 등을 자연스럽게 표현하는 기술이 요구되고 있다.

본 논문에서는 새로운 화자의 소량의 음성 데이터를 이용하여 해당 화자의 음색으로 합성음을 생성할 수 있는 다화자 음성합성 기술에 중점을 둔다. 이는 소량의 음성 데이터로 학습에 사용되지 않은 화자의 음색을 표현해야 하는 난이도가 높은 과제이다. 화자의 음색 특징을 모델링하기 위해서 화자 임베딩에 대한 연구가 진행되며, 이는 화자에 따라 다른 아이덴티티를 부여하거나 화자의 음향 특징을 적절히 압축하는 형태로 이뤄진다. 화자의 특성을 나타내는 가존 방식은 음향 특징 파라미터를 이용하는데 반해, 본 논문에서는 종단형 방식으로 화자의 아이덴티티를 부여하는 형태로 화자 임베딩에 대한 연구를 진행한다.

본 논문에서는 화자인식 기술에서 사용하는 딥러닝 기반 x-vector 기법을 적용하여 화자 인코더를 구성하고, 화자 인코더를 통해 소량의 데이터로 새로운 화자의 음색을 합성하는 기술을 제안한다.

본 논문에서 2장에서는 x-vector를 이용한 화자 임베딩 기술을, 3장에서는 실험 환경과 실험에 사용하는 음성 DB에 대해 기술한다. 4장에서는 제안하는 화자 임베딩과 다화자 음성합성 모델을 학습하는 방법에 대해 설명한다. 5장에서는 실험 결과 및 성능에 대해 기술하며, 마지막으로 6장에서 결론을 맺는다.

Ⅱ. 화자 임베딩을 위한 x-vector 기법

화자 임베딩 기술은 매우 많은 화자의 음성 DB로 화자의 특징을 학습하여 화자의 음색을 구분할 수 있는 특징인 발화 속도와 억양 등을 모델링한다. 그리고 이를 이용하여 학습에 포함되지 않은 화자의 적은 음성 데이터로부터 화자의 특징을 나타내는 임베딩 벡터를 생성하는 방식이다. 화자 임베딩 기술에는 대표적으로 JFA(Joint Factor Analysis), GMM(Gaussian Mixture Model), i-vector 등이 있으며, 딥러닝 기반의 화자 특징 추출 기술로는 d-vector와 x-vector 방식이 있다[3].

d-vector와 x-vector 방식은 화자 인식 신경망 계층 중에서 한 개의 은닉층을 임베딩으로 이용한다는 점이 유사하나, 발화 레벨 임베딩을 취하는 방식에서 차이가 있다. d-vector는 전연결 심층신경망(Fully-connected Deep Neural Networks)의 마지막 은닉층을 특징으로 이용하고, x-vector는 시간 지연 구조 신경망의 마지막 은닉층을 통계적으로 풀링하여 사용한다[3]. 본 논문에서는 화자 임베딩을 위해 x-vector를 이용한다.

그림 1은 화자 인식 신경망 구조와 x-vector를 이용

한 화자 임베딩 방식을 보여준다[4]. 음향 특징 데이터 $\{x_1, x_2, ..., x_T\}$를 입력으로 받아들이는 계층 l_1에서 l_5까지 5개의 계층은 시간 지연 구조로 구성되어 있으며 프레임 수준에서 작동한다. 현재 시간을 t라고 할 때, 입력 계층 l_1에서 프레임이 $\{t-2, t-1, t, t+1, t+2\}$로 분할된다. 다음 l_2와 l_3 계층은 시간 단계 $\{t-2, t, t+2\}$와 $\{t-3, t, t+3\}$에서 이전 계층의 출력을 분할한다. 계층 l_4와 l_5에서는 시간적 문맥이 추가되지 않으며, 따라서 처음 5개의 계층은 15개 프레임의 시간적 문맥 정보를 처리하게 된다. 통계적 풀링 계층에서는 신경망의 프레임 수준 출력 벡터를 입력으로 받아 평균과 표준편차를 계산한다. 이러한 풀링 과정은 다양한 길이의 음성 세그먼트로부터 고정된 길이의 특징 벡터를 만들게 된다. 평균과 표준편차는 두 개의 은닉층 l_6와 l_7, 그리고 최종적으로 소프트맥스 출력 계층에 전달되어 화자를 분류하게 된다. 학습을 마친 후, x-vector 방식에서 화자 임베딩은 l_6 계층의 출력에서 도출된다[4][5].

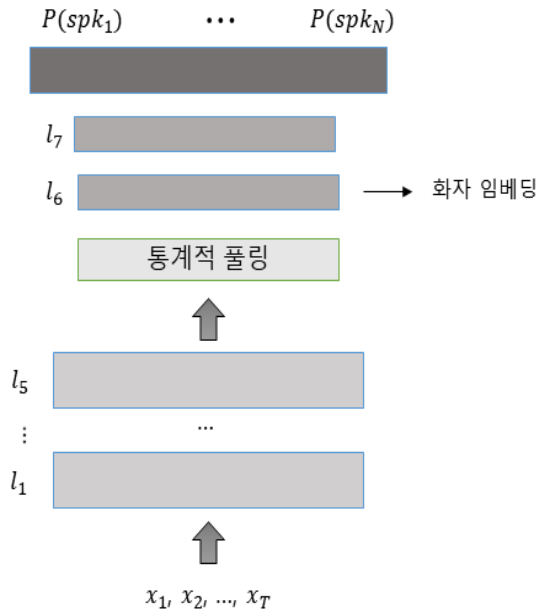

그림 1. 화자 인식 신경망 구조와 화자 임베딩
Figure 1. Structure of the speaker recognition neural networks and speaker embedding

III. 실험 환경 및 음성 DB

1. 실험 환경

x-vector 모델과 다화자 음성합성 모델을 학습하는 컴퓨터 환경은, 우분투 16.04LTS 운영체제, 인텔 i7-8700K CPU, 64GB 메인 메모리, 그리고 Nvidia GTX 1080ti GPU 2개 등이다. 소프트웨어는 파이썬 버전 3.6.9, 파이토치 버전 1.5.0, CUDA 버전 10.1, Kaldi와 ESPnet-TTS 등이다.

2. 음성 DB

1) x-vector 모델 학습용 음성 데이터

x-vector 모델 학습 데이터는 영어 데이터 셋과 한국어 데이터 셋 등 두 가지 데이터 셋으로 구성되어 있다.

영어 음성 DB는 전 세계 유명 인사들이 발화한 음성 데이터를 수집한 Voxceleb 데이터 셋을 사용한다[6]. 이 데이터 셋은 음성 기술 분야에서 널리 사용되고 있으며, Voxceleb1과 2로 구성되어 있다. 약 7,000명의 화자가 발화한 약 130만개의 음성 파일로 이루어져 있으며, 남녀 비율은 남성 화자 39%, 여성 화자 61%,이다.

한국어 음성 DB는 OpenSLR에서 제공하는 한국어 데이터인 Zeroth-Korean[7]과, 네이버 클로바 AI에서 제공하는 대화체 음성 데이터[8] 등을 이용한다. 한국어 음성 DB는 약 8,900명의 화자가 발화한 약 8만 여개의 음성 파일로 구성되어 있으며, 남녀 비율은 남성 35%, 여성 65%,이다.

이들 음성 DB는 음성 파일과 메타 데이터로 구성되어 있고, 메타 데이터에는 화자 아이디, 성별, 국적 정보를 포함하고 있다. x-vector 모델을 학습하기 위해 필요한 데이터는 화자 아이디와 이 화자가 발화한 음성 파일이다. 음성 데이터 형식은 선형 PCM 16 비트와 표본화 주파수 16kHz로 통일하여 사용한다.

2) 디화자 음성합성 모델 학습용 음성 데이터

다화자 음성합성 학습 데이터도 마찬가지로 영어 데이터 셋과 한국어 데이터 셋으로 구성되어 있다.

영어 음성 데이터는 다화자 음성합성 연구에서 널리 쓰이는 LibriTTS[9]를 사용한다. 이 음성 DB는 약 2,700명의 화자가 발화한 약 16만개의 음성 파일로 분량은 약 585시간으로 구성되어 있으며, 남녀 비율은 5:5로 균등하게 분배되어 있다.

한국어 음성 데이터는 자체적으로 보유하고 있는 음성 DB를 이용한다. 이 데이터는 66명의 화자가 발화한 약

3.4만 여개의 음성 파일로 구성되어 있으며, 남녀 비율은 남성 35%, 여성 65%,이다.

이들 음성 DB에는, 음성합성 모델을 학습하기 위해 필요한 음성 파일과 각 음성 파일의 텍스트 정보가 포함되어 있다. 음성 데이터 형식은 선형 PCM 16 비트와 표본화 주파수 24kHz로 통일하여 사용한다.

3) x-vector 성능 평가용 음성 데이터

성능 평가용 데이터도 마찬가지로 영어와 한국어 데이터 셋 등 두 가지 데이터 셋으로 구성되어 있다. 평가용 데이터이므로 학습용 데이터에서처럼 메타 데이터가 필요하지 하지 않고 음성 데이터만 사용한다.

영어 데이터는 화자인식과 음성인식 성능 평가에서 널리 사용하는 SITW 데이터[10]를 이용한다. 이 데이터는 약 2,800명의 화자가 발화한 음성 파일로 구성되어 있다.

한국어 데이터는 자체적으로 보유하고 있는 데이터를 사용하는데, 이 데이터는 4명의 화자가 발화한 800개의 음성 파일로 구성되어 있다.

Ⅳ. 실험 방법

이 절에서는 제안하는 x-vector 모델과 다화자 음성합성 모델을 학습하는 방법에 대하여 설명한다.

본 논문에서는 Kaldi와 ESPnet-TTS 툴킷으로 화자 임베딩을 이용한 다화자 음성합성 시스템을 구현한다. Kaldi[11]는 음성 인식 시스템을 구현하는데 필요한 라이브러리를 제공하는 오픈 소스 툴킷이다. ESPnet-TTS[12]는 종단형 음성합성 시스템을 구현할 수 있는 오픈 소스 툴킷이며, 음성 인식과 음성 처리를 위한 오픈 소스인 ESPnet[13]의 확장이다. ESPnet-TTS는 Kaldi 스타일에 맞춰 데이터 및 음성 처리 기법을 제공하므로, 이 두 개의 툴킷은 데이터 처리 과정이 호환된다. 화자 임베딩을 위한 x-vector 생성은 Kaldi에서 제공하는 레시피를, 다화자 음성합성 시스템을 구현하기 위해서는 EPnet-TTS에서 제공하는 LibriTTS 레시피를 활용한다.

1. x-vector 모델 학습

x-vector 모델을 학습하는 과정은 다음과 같다. 주어진 음성 데이터에서 묵음을 제거하기 위해 끝점 추출을 하여 음성 부분만 잘라내고, x-vector 모델에서 사용하는 특징 파라미터인 MFCC(Mel-Frequency Cepstral Coefficients)를 추출한다. 또한 x-vector 모델의 강인성을 더하기 위해 Reverberation, Babble 잡음 및 음악 등을 음성 데이터에 추가하여 데이터를 증강한다. 여기에서 사용하는 잡음은 OpenSLR에서 제공하는 Reverberation 잡음 데이터[14]와 Musan 데이터[15]를 이용한다. 증강된 데이터에서도 MFCC를 추출하여, 잡음이 없는 음성 데이터에서 추출한 MFCC와 결합하여 x-vector 모델의 학습에 사용한다.

그림 1의 화자 임베딩 신경망을 학습하면 소프트맥스 출력 계층에서 화자 인식 결과를, 계층 l_6에서 x-vector를 도출할 수 있다. 학습을 마치고 계층 l_6에서 추출한 화자 임베딩을 위한 x-vector를 다화자 음성합성 모델에서 사용한다.

2. 다화자 음성합성 모델 학습

그림 2는 다화자 음성합성 모델의 구조를 나타낸다. 텍스트 입력에서 멜-스펙트로그램을 합성하는 모듈은 Tacotron2[16]로 구성되어 있고, 멜-스펙트로그램에서 합성음을 만드는 보코더는 16 비트 음성 샘플을 생성하기 위해 로지스틱 혼합 분포(Mixture of Logistic distributions, MoL)가 적용된 WaveNet[16][17]이다. 다화자 음성합성을 위해 Tacotron2에 화자 임베딩 모듈이 추가되어 있다.

준비된 다화자 음성합성 학습 데이터에서, 먼저 표본화 주파수를 24kHz에서 16kHz로 변환하여 30 차원의 MFCC를 추출한다. 그리고 앞 절에서 학습한 x-vector 모델을 이용하여 음성 파일별로 화자 임베딩 벡터를 생성한다. 양방향 LSTM(Bidirectional Long Short-Term Memory)의 출력과 화자 임베딩 벡터를 연결하여(Concatenate) 그림 2의 다화자 음성합성 모델을 학습한다. 다화자 음성합성 모델을 학습하기 위해 필요한 입력 데이터는 음성 파일과 각 음성 파일에 해당하는 텍스트 데이터와 x-vector이다.

성능 평가를 위해 합성음을 생성할 때 필요한 입력 데이터는 원하는 텍스트 데이터와 수 초 길이의 음성 파일이다. 이 음성 파일로부터 화자 임베딩을 통해 이 화자의 음색을 표현한다.

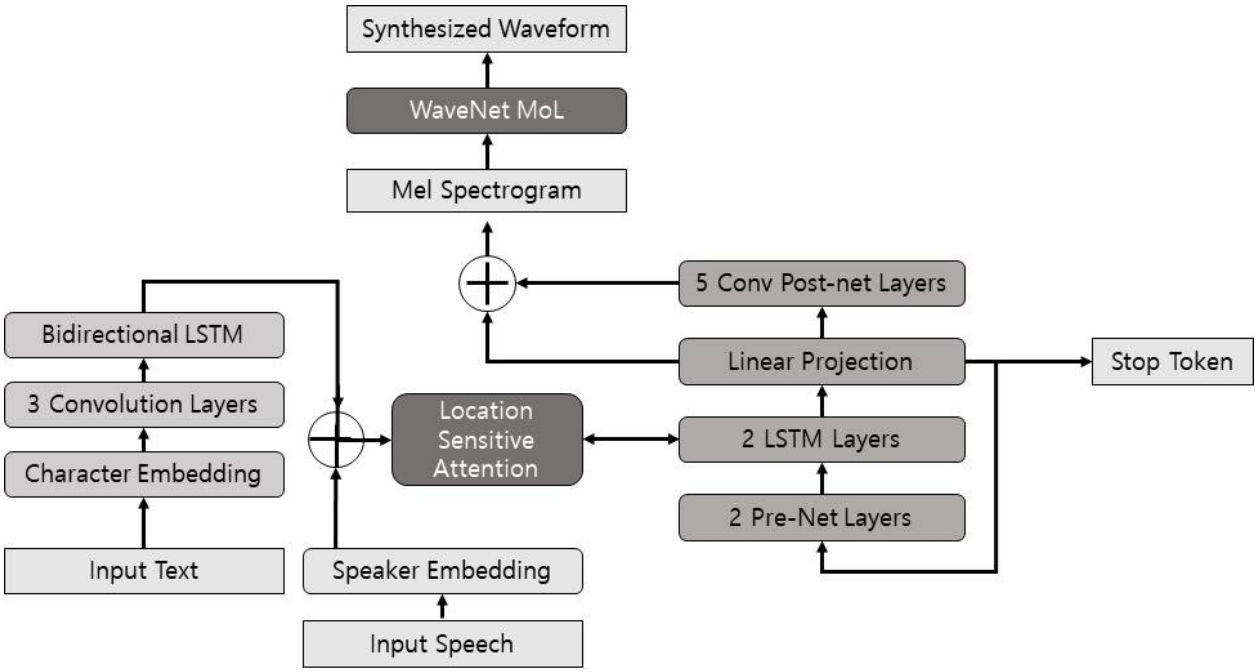

그림 2. 다화자 음성합성 모델 구조
Figure 2. Structure of the multi-speaker speech synthesis model

V. 실험 결과

x-vector 기법의 성능은 동일오류율(Equal Error Rate, EER)을 이용하여 평가한다. EER은 화자인식 분야에서 주로 사용하는 성능 지표로, 오인식률(False Acceptance Rate)과 오거부율(False Rejection Rate)이 같아지는 비율이다.

표 1은 논문 [5]에서 진행한 i-vector와 x-vector 연구의 실험 결과와, 본 논문에서 실험한 결과의 성능을 비교한 표이다. 먼저 평가용 음성 데이터에서 x-vector를 추출하여 평균을 계산하고, LDA(Linear Discriminant Analysis)를 이용하여 128 차원으로 감소시킨 뒤 PLDA (Probabilistic LDA) 모델을 학습하여 성능 평가에 사용한다.

영어 데이터를 이용하여 성능을 평가한 결과에서 EER은 3.55%이고, 한국어 데이터에서는 18.79%,이다. 영어 데이터에서는 논문 [5]의 성능 보다 우수하나, 한국어 데이터에서는 성능이 좋지 않다. 이것은 학습에 사용된 한국어 음성 데이터의 수가 부족하기 때문이라고 판단된다.

다화자 음성합성 시스템의 성능은 어텐션(Attention) 그래프와, 원 음성과 합성음의 멜-스펙트로그램을 비교하여 평가한다. 어텐션은 텍스트와 음성신호 간의 정렬(Alignment)을 나타낸다. 학습이 잘 된 경우의 어텐션은 그림 3, 4에서와 같이 가느다란 대각선의 형태를 보여준다. 그림 3은 영어 음성합성, 그림 4는 한국어 음성합성의 결과를 보여준다.

표 1. x-vector 기법의 성능 평가 결과
Table 1. Performance results of the x-vector technique

	EER
논문 [5] i-vector	7.45%
논문 [5] x-vector	4.16%
영어 데이터 x-vector	3.55%
한국어 데이터 x-vector	18.79%

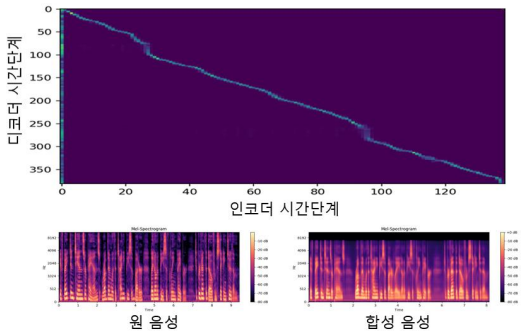

그림 3. 영어 음성합성에서 어텐션 그래프(위)와 멜-스펙트로그램(아래)
Figure 3. Attention graph(above) and mel-spectrogram (bottom) in English speech synthesis

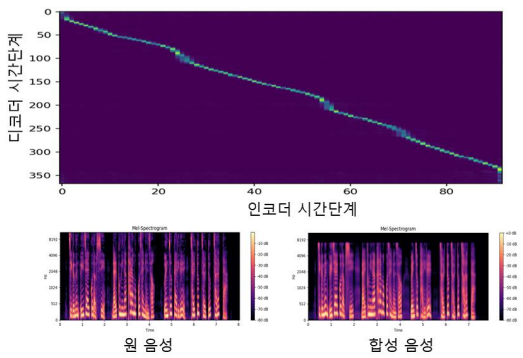

그림 4. 한국어 음성합성에서 어텐션 그래프(위)와 멜-스펙트로그램(아래)
Figure 4. Attention graph(above) and mel-spectrogram (bottom) in Korean speech synthesis

다화자 음성합성 시스템의 합성음에 대해 수행한 청취 평가 결과는, 합성 음성의 품질이 원 음성과 차이가 있으나, 영어와 한국어 모두 자연스러운 합성음을 생성한다고 평가되었다.

Ⅵ. 결 론

딥러닝 기반 다화자 TTS 시스템에서 자연스러운 합성음을 만들기 위해서는 많은 화자가 발화한 대용량의 고품질 음성 DB로 학습하는 것이 필요하다. 그러나 이러한 음성 DB를 수집하는 것은 매우 어려운 작업이므로, 매우 많은 화자의 소량의 음성 DB를 사용하여 학습하고, 이로부터 다화자의 음성을 생성해 내는 기술이 필요하다. 본 논문은 특정 화자의 적은 음성 데이터를 이용하여 해당 화자의 합성음을 생성하는 다화자 음성합성 시스템에 대한 연구이다.

본 논문에서는 특정 화자 목소리의 특징을 추출하기 위해 화자 임베딩 기술인 x-vector 모델을 학습하고, 화자별로 x-vector를 추출하여 멜-스펙트로그램 생성 기술인 Tacotron2 모델 학습에 입력으로 추가하고, Wavenet 보코더를 이용하여 합성음을 생성하는 다화자 음성합성 시스템을 구성하였다.

화자의 특징을 추출하는 x-vector는 화자인식 기술에서 사용하는 평가 지표인 EER을 이용하여 성능을 평가하였고, 다화자 음성합성 시스템은 어텐션과 멜-스펙트로그램으로 성능을 확인하였다. 영어 데이터에 대해 x-vector 기법을 실험한 결과는 기존 연구 대비 약 14.7%의 성능이 향상되었으나, 한국어 데이터인 경우 성능이 나빠졌다. 이것은 x-vector 모델 학습에 사용된 한국어 데이터가 충분하지 않기 때문이라고 판단된다.

향후 연구로는 한국어 데이터에 적합한 TTS 모델을 연구하고, 소량의 음성 데이터를 이용하여 성능이 향상된 한국어 다화자 음성 합성 시스템에 대한 연구를 진행할 계획이다.

References

[1] C. H. Kwon, “Performance comparison of state-of-the-art vocoder technology based on deep learning in a Korean TTS system", *The Journal of the Convergence on Culture Technology (JCCT)*, Vol. 6, No. 2, pp. 509-514, 2020, DOI:10.17703/JCCT.2020.6.2.509

[2] C. H. Kwon, “Comparison of Korean real-time text-to-speech technology based on deep learning”, *The Journal of the Convergence on Culture Technology (JCCT)*, Vol. 7, No. 1, pp. 640-645, 2021, DOI:10.17703/JCCT.2021.7.1.640

[3] M. S. Jo, “A study on a multi-speaker TTS system using speaker embedding”, *Master Thesis*, Graduate School of Daejeon Univ. 2021

[4] D. Snyder, D. Garcia-Romero, D. Povey, S. Khudanpur, “Deep neural network embeddings for text-independent speaker verification”, *Proceedings of the Interspeech 2017*, pp. 999-1003, 2017, DOI:10.21437/Interspeech.2017-620

[5] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, “X-vectors: Robust DNN embeddings for speaker recognition”, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2018*, pp. 5329-5333, 2018, DOI:10.1109/ICASSP.2018.8461375

[6] A. Nagrani, J. S. Chung, A. Zisserman, “VoxCeleb: A large-scale speaker identification dataset”, *Proceedings of the Interspeech 2017*, pp. 2616-2620, 2017, DOI:10.1109/ICASSP.2018.8461375

[7] Zeroth-Korean: Korean open source speech corpus, https://www.openslr.org/40/

[8] J. W. Ha, K. H. Nam, J. Kang, et al., “ClovaCall: Korean goal-oriented dialog speech corpus for automatic speech recognition of contact centers”, *Proceedings of the Interspeech 2020*, pp. 409-413, 2020, DOI:10.21437/Interspeech.2020-1136

[9] H. Zen, V. Dang, R. Clark, et al., “LibriTTS: A corpus derived from LibriSpeech for text-to-speech”, *Proceedings of the Interspeech 2019*, pp. 1526-1530,

2019, DOI:10.21437/Interspeech.2019-2441
[10]M. McLaren, L. Ferrer, D. Castan, A. Lawson, "The speakers in the wild (SITW) speaker recognition database", *Proceedings of the Interspeech 2016*, pp. 818-822, 2016, DOI:10.21437/Interspeech.2016-1129
[11]D. Povey, A. Ghoshal, G. Boulianne, et al., "The Kaldi speech recognition toolkit", *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) 2011*, 2011
[12]T. Hayashi, R, Yamamoto, K. Inoue, et al., "ESP net-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2020*, pp. 7654-7658, 2020, DOI:10.1109/ICASSP40776.2020.9053512
[13]S. Watanabe, T. Hori, S. Karita, et al., "ESPnet: End-to-end speech processing toolkit", *Proceedings of the Interspeech 2018*, pp. 2207-2211, 2018, DOI:10.21437/Interspeech.2018-1456
[14]T. Ko, V. Peddinti, D. Povey, M. Seltzer, S. Khudanpur, "A study on data augmentation of reverberant of speech for robust speech recognition", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2017*, pp. 5220-5224, 2017, DOI:10.1109/ICASSP.2017.7953152
[15]D. Snyder, G. Chen, D. Povey, "MUSAN: A music, speech, and noise corpus", *arXiv preprint*. https://arxiv.org/pdf/1510.08484.pdf, 2015 Oct.
[16]J. Shen, R. Pang, R. J. Weiss, et al., "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2018*, pp. 4779-4783, 2018, DOI: 10.1109/ICASSP.2018.8461368
[17]A. Oord, S. Dieleman, H. Zen, et al., "WaveNet: A generative model for raw audio", *Proceedings of the 9th ISCA Speech Synthesis Workshop*, pp. 125-125., 2016

※ 이 논문은 한국연구재단 지역대학 우수과학자 지원 사업(NRF-2020R1I1A3052136)에 의해 연구되었음