한국음향학회지 제41권 제3호 pp. 359~366 (2022)
The Journal of the Acoustical Society of Korea Vol.41, No.3 (2022)
https://doi.org/10.7776/ASK.2022.41.3.359

pISSN : 1225-4428
eISSN : 2287-3775

콘포머 기반 FastSpeech2를 이용한 한국어 음식 주문 문장 음성합성기

A Korean menu-ordering sentence text-to-speech system using conformer-based FastSpeech2

최예린,[1] 장재후,[1] 구명완[1†]

(Yerin Choi,[1] JaeHoo Jang,[1] and Myoung-Wan Koo[1†])

[1]서강대학교 인공지능학과 대학원 지능형 음성대화 인터페이스 연구실

(Received March 21, 2022; revised April 29, 2022; accepted May 23, 2022)

초 록: 본 논문에서는 콘포머 기반 FastSpeech2를 이용한 한국어 메뉴 음성합성기를 제안한다. 콘포머는 본래 음성인식 분야에서 제안된 것으로, 합성곱 신경망과 트랜스포머를 결합하여 광역과 지역 정보를 모두 잘 추출할 수 있도록 한 구조다. 이를 위해 순방향 신경망을 반으로 나누어 제일 처음과 마지막에 위치시켜 멀티 헤드 셀프 어텐션 모듈과 합성곱 신경망을 감싸는 마카론 구조를 구성했다. 본 연구에서는 한국어 음성인식에서 좋은 성능이 확인된 콘포머 구조를 한국어 음성합성에 도입하였다. 기존 음성합성 모델과의 비교를 위하여 트랜스포머 기반의 FastSpeech2와 콘포머 기반의 FastSpeech2를 학습하였다. 이때 데이터셋은 음소 분포를 고려한 자체 제작 데이터셋을 이용하였다. 특히 일반대화 뿐만 아니라, 음식 주문 문장 특화 코퍼스를 제작하고 이를 음성합성 훈련에 사용하였다. 이를 통해 외래어 발음에 대한 기존 음성합성 시스템의 문제점을 보완하였다. ParallelWave GAN을 이용하여 합성음을 생성하고 평가한 결과, 콘포머 기반의 FastSpeech2가 월등한 성능인 MOS 4.04을 달성했다. 본 연구를 통해 한국어 음성합성 모델에서, 동일한 구조를 트랜스포머에서 콘포머로 변경하였을 때 성능이 개선됨을 확인하였다.
핵심용어: Text-to-Speech (TTS), 음성합성, 콘포머, 딥러닝

ABSTRACT: In this paper, we present the Korean menu-ordering Sentence Text-to-Speech (TTS) system using conformer-based FastSpeech2. Conformer is the convolution-augmented transformer, which was originally proposed in Speech Recognition. Combining two different structures, the Conformer extracts better local and global features . It comprises two half Feed Forward module at the front and the end, sandwiching the Multi-Head Self-Attention module and Convolution module. We introduce the Conformer in Korean TTS, as we know it works well in Korean Speech Recognition. For comparison between transformer-based TTS model and Conformer-based one, we train FastSpeech2 and Conformer-based FastSpeech2. We collected a phoneme-balanced data set and used this for training our models. This corpus comprises not only general conversation, but also menu-ordering conversation consisting mainly of loanwords. This data set is the solution to the current Korean TTS model's degradation in loanwords. As a result of generating a synthesized sound using ParallelWave Gan, the Conformer-based FastSpeech2 achieved superior performance of MOS 4.04. We confirm that the model performance improved when the same structure was changed from transformer to Conformer in the Korean TTS.
Keywords: Text-to-Speech (TTS), Speech synthesis, Conformer, Deep learning

PACS numbers: 43.72.Ja, 43.72.Kb

†**Corresponding author:** Myoung-Wan Koo (mwkoo@sogang.ac.kr)
AI Graduate School, Sogang University, 35, Baekbeom-ro, Mapo-gu, Seoul 04107, Republic of Korea
(Tel: 82-2-705-8935, Fax: 82-2-704-8273)

I. 서 론

음성합성(Text-to-Speech, TTS) 시스템은 텍스트를 입력으로 받아 음성을 출력하는 시스템이다. TTS 시스템 개발을 위해서 음편 선정 방식[1]이나, 은닉 마르코프 모형(Hidden Markow model, HMM) 기반 통계적 방식[2] 등 다양한 방법론들이 이전부터 적용되어 왔으며, 현재 가장 뛰어난 성능을 보이는 것은 딥러닝 기반 TTS 방식이다. 딥러닝 기반 TTS는 음성합성을 위해서 텍스트를 입력으로 준 뒤, 멜 스펙트로그램을 예측하고 멜 스펙트로그램으로 음성파일을 생성하는 2가지 단계를 거친다. 이때 멜 스펙트로그램을 예측하는 모델을 Text2Mel 모델이라 하며 Tactron2,[3] FastSpeech2[4] 등이 있다. 멜 스펙트로그램을 입력으로 사용하여 음성을 생성하는 모델은 보코더 모델이라 하며, WaveNet,[5] ParallelWave GAN[6] 등이 있다.

초기 딥러닝 기반 TTS 방법론들은 자기회귀방식을 주로 사용하였는데, 이 방식은 추론 속도가 느려 실시간 처리가 불가능하다는 단점이 존재했다. 하지만 병렬적으로 멜 스팩토그램을 생성하는 비자기회귀적 TTS 모델들은 자기회귀 방식의 딥러닝 기반 TTS 모델이 합성하는 음성 품질에 맞먹는 성능을 가지면서도 훨씬 빠른 추론 속도를 보이며 단점을 극복하였다. 근래에 제안된 비자기회귀적 TTS인 FastSpeech[7]의 경우 이전 SOTA 모델보다 270배 빠른 멜 스펙트로그램 생성속도와 38배 빠른 추론 속도를 달성하여 딥러닝 기반 TTS 기술을 활용한 실시간 음성 합성이 가능함을 시사했다.[7]

본 연구에서는 최근 음성인식 분야에서 성능적 우위를 보이는 모델인 Conformer(이하 콘포머) 구조 기반의 음성합성기를 제안한다.[8] 콘포머는 트랜스포머의 인코더에 합성곱 신경망(Convolutional Neural Network, CNN)을 결합하여 변형시킨 모델이며 트랜스포머 기반 한국어 음성인식 모델 보다 뛰어난 성능을 보인다.[9] 음성 합성 분야 에서도 기존 음성 합성 모델에 콘포머를 도입하는 연구가 진행 되고 있으며, 트랜스포머 기반 음성합성 모델과의 비교를 통해 음성 품질의 향상이 확인되었다.[10]

또한 외래어에 대해서도 일정한 성능을 가진 TTS 시스템을 만들기 위해서, 고유어 기반 한국어 문장 코퍼스에 외래어 희소 음소열로 표기된 음식 주문 문장 코퍼스를 첨가하여 학습 데이터셋을 구축하였다. 현재 국내에서 개발된 딥러닝 기반 TTS 시스템들은 대부분이 고유어 기반으로 학습되었기 때문에, 외래어가 다수 담긴 음식 주문 문장에 대해서는 합성음 품질이 낮다. 또한 음식 주문 문장에 최적화된 음성 합성기 개발 사례는 존재하지 않는다. 위와 같은 문제를 해결하기 위해서, 본 연구에서는 고유어 및 외래어 음소열의 분포 균형을 고려하여 구축한 데이터셋으로 TTS 모델 학습을 진행하였다. 그 결과 학습한 TTS 모델은 고유어뿐만 아니라 외래어에 대해서도 일정하게 높은 품질의 합성음을 생성하였다.

최종적으로, 본 연구에서는 콘포머 구조 기반의 한국어 음식 주문 문장 TTS 시스템을 제안한다. 2장에서는 콘포머와 트랜스포머를 비교하고, 3장에서는 콘포머 기반 음성합성 모델에 대해 설명한다. 4장에서는 훈련에 사용한 데이터셋에 대해 기술하고 훈련을 결과를 확인하며, 5장에서는 결론과 논의로 글을 맺는다.

II. 콘포머와 트랜스포머 비교

2.1 Transformer

트랜스포머는[11] 기존 RNN 기반 seq2seq 구조의 고질적인 정보 손실 문제와 느린 연산속도 문제를 완화한 기계 번역 모델이다. 입력 텍스트에는 위치 인코딩을 적용하여 텍스트 임베딩 행렬에 각 단어의 위치 정보를 직접 반영한다. 이는 순차적으로 은닉 벡터 계산이 이루어지는 RNN 기반 층(layer)을 필요로 하지 않아 모델의 연산속도를 높이는 데 기여한다.

모델의 구조는 크게 자기 집중 모듈과 순방향 신경망으로 구성된 인코더 및 디코더로 이루어져 있다. 구조도는 Fig. 1과 같다. 트랜스포머 모델에서 사용하는 어텐션 층들은 멀티 헤드 어텐션으로, 전체 입력에 대해 병렬적으로 어텐션을 수행하여 Q(query), K(key), V(value)를 산출하는 방식이다. 멀티헤드 어텐션은 하이퍼 파라미터인 헤드 수만큼 Q, K, V 각각의 차원을 나누어 어텐션을 수행하는 방법으로 입력 텍스트를 여러 시점으로 바라보면서 어텐션 값을 계

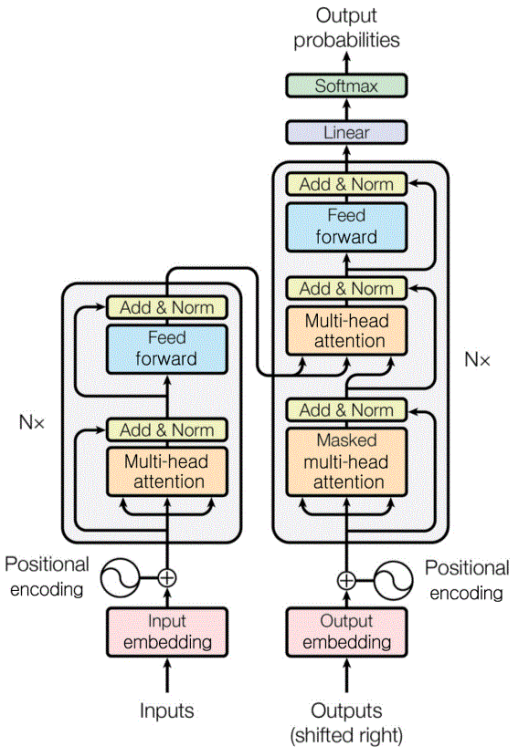

Fig. 1. (Color available online) Model architecture of transformer.

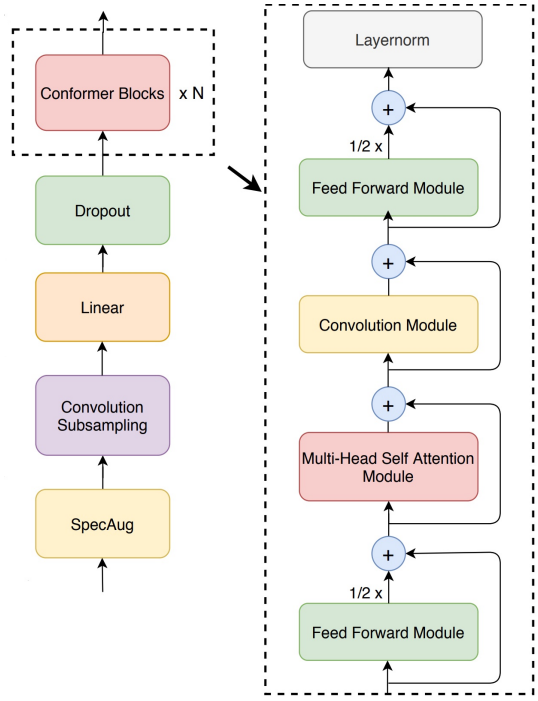

Fig. 2. (Color available online) Encoder architecture of conformer.

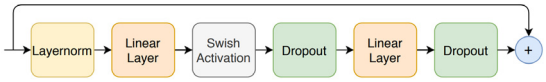

Fig. 3. (Color available online) Feed forward network of conformer.

산한다. 또한 일반적인 RNN 기반 번역 모델에서 사용하는 어텐션 방식과의 차이점이 있다. 멀티 헤드 어텐션 방식은 어텐션 값 계산 시 입력 문장 내 단어 사이의 연관도와 출력 간 무게 벡터를 함께 반영한다. 이는 기존의 번역 모델에 비해서 뛰어난 성능을 가지게 한다.

음성 합성의 경우, 인코더에서는 입력 텍스트를 받아 Q, V을 추출하고 이는 디코더의 두번째 하위층에 존재하는 멀티 헤드 어텐션에 사용된다. 디코더에서는 음향 정보인 멜 스펙트로그램을 예측한다.

2.2 Conformer

콘포머[8]는 이름에서도 알 수 있듯이 합성곱 신경망과 트랜스포머를 합친 구조다. 지역 정보를 잘 잡아내는 합성곱 신경망과 광역 정보에 대한 성능이 좋은 트랜스포머를 합쳐, 지역과 광역 정보 모두에 대해 좋은 성능을 낼 수 있는 구조를 제안하였다. 콘포머 구조는 트랜스포머와 마찬가지로 인코더-디코더로 구성된다. 여기서 디코더는 트랜스포머의 디코더와 동일하다. 콘포머가 트랜스포머와 차별점을 가지는 부분은 인코더 부분이다. Fig. 2에서 콘포머의 인코더 구조를 확인할 수 있다. 트랜스포머와의 비교는 다음과 같다.

트랜스포머는 멀티 헤드 셀프 어텐션 모듈과 층 정규화 다음에 순방향 신경망을 거친다(Fig. 1). 반면에 Fig. 2 오른쪽 구조도와 같이 콘포머는 순방향 신경망을 반으로 나누어 제일 처음과 마지막에 지나도록 하였다. 이렇게 반으로 나누어 전체 구조를 감싸게 한 것을 마카론 구조라 한다. Fig. 3에서 콘포머에서 사용한 순방향 신경망 구조를 확인할 수 있다. 트랜스포머와 달리 순방향 신경망 안에서 Rectified Linear Unit(ReLU)가 아닌 Swish 활성화 함수를 사용하여 더 좋은 성능을 내었다.[12]

Fig. 4에 나타난 멀티 헤드 셀프 어텐션 모듈에서는 상대 위치 임베딩을 활용하여 입력 길이가 달라

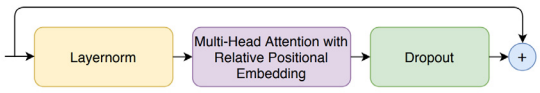

Fig. 4. (Color available online) Multi-head self-attention module of conformer.

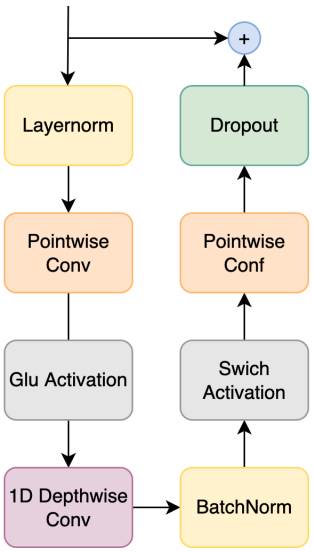

Fig. 5. (Color available online) Convolution module of conformer.

져도 성능을 유지할 수 있도록 하였다. 또한 트랜스포머와 달리 층 정규화를 제일 먼저 수행하여 깊은 모델 학습을 원활하게 하였다.

합성곱 신경망 모듈(Fig. 5)에서는 점별 합성곱과 게이트 선형 유닛(Gated Linear Units, GLU)을 지난다. 그 후, 1차원 채널별 합성곱과 배치 정규화를 한 후, Swish 함수와 점별 합성곱을 차례로 지난다. 마지막으로 과적합을 방지하기 위한 드롭아웃을 수행한다.

콘포머는 트랜스포머와 달리 신경망을 반으로 나누어 처음과 마지막에 배치했고, 이때 활성화 함수를 ReLU에서 Swish 함수로 변경하였다. 멀티 헤드 셀프 어텐션 모듈에서는 층 정규화 과정의 위치를 변경하여 깊은 모델 학습에 대한 성능을 높였다. 이후, 합성곱 모듈을 추가하였다.

III. 콘포머를 이용한 한국어 음성합성

3.1 콘포머 기반 FastSpeech2

FastSpeech2[2]의 선행 기술인 Fastspeech[7]는 기존의 자기회귀 방식을 사용하는 음성합성 모델들의 단점을 보완한 모델이다. 기존 모델들은 텍스트가 길어지면 성능이 저하되는 문제뿐만 아니라 추론 속도가 느리다는 단점을 가지고 있었다. FastSpeech는 자기회귀 방식이 아닌 병렬적으로 시퀀스를 생성하는 방식을 사용하여 이를 극복하였다. 해당 방식은 이전 결과에 의존적이지 않기 때문에 추론 속도가 빠르다. 또한 FastSpeech는 내부 구조에 길이 조절기를 추가하여 입력(텍스트)과 출력(음성)의 길이 차이에도 일정한 성능을 보였다.

FastSpeech2[2]는 FastSpeech의 교사-학생 학습 구조를 단순화한 모델이다. 더 나아가 구조는 더 단순해졌지만 성능은 떨어지지 않는 성과를 냈다. FastSpeech와 달리 교사 모델 학습 과정이 생략되어 학습에 걸리는 시간을 단축할 수 있었다. 또한, 음고와 에너지 등을 처리하는 모듈을 추가하여 더 다양한 음성정보를 표현할 수 있게 하였다.

콘포머 기반 FastSpeech2는 기존 FastSpeech2의 트랜스포머를 콘포머로 바꾼 구조다.

IV. 실험 및 결과

4.1 데이터셋

본 연구에서는 단일 화자 한국인 여성 성우 목소리 기반의 음성 코퍼스를 자체 제작하여 활용하였다. 해당 데이터셋은 일상적 대화 상황에서의 발화 내용이 담긴 고유어 위주로 구성된 10,000문장과, 고객이 음식점에서 메뉴를 주문하는 외래어 위주의 발화 내용이 담긴 3,000문장을 합하여 총 13,000문장(26.8시간)의 발화 음성 파일 및 이에 상응하는 텍스트 파일로 이루어져 있다. 즉, 일반 대화 문장이 주를 이루고 음식 주문 문장을 첨가하여 코퍼스를 구축한 형식이다. 이때 일반 대화 문장의 텍스트 데이터는 AI hub의 일반 대화 데이터[1]에서 추출하였으며, 음식 주문 문장 텍스트 데이터는 자체 제작 하였다. 범용적인 일반 대화 문장뿐만 아니라 외래어 위주의 음식 주문 문장에 대해서도 잘 작동하는 음성합성기

1) AI hub 한국어 음성 data: http s://aihub.or.kr/aidata/105/download

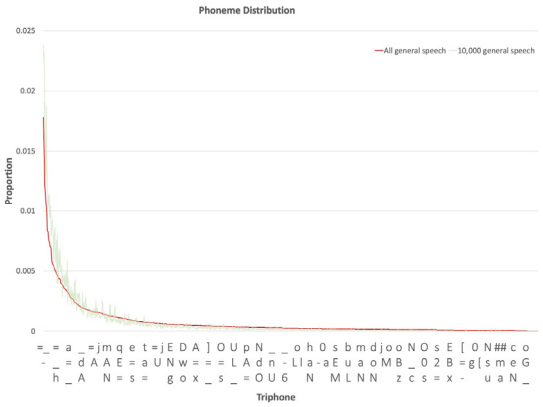

Fig. 6. (Color available online) Triphone distribution of general speech.

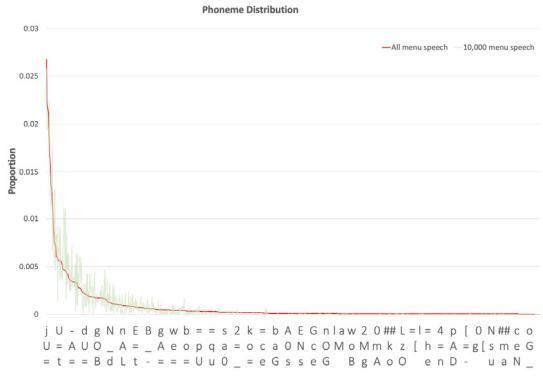

Fig. 7. (Color available online) Triphone distribution of menu-ordering speech.

훈련을 위해 이러한 코퍼스를 구축하였다.

4.1.1 음소 분포를 활용한 데이터셋 최적화

음성합성 모델을 훈련할 때 입력값은 자소가 아닌 음소를 사용한다. 따라서 데이터셋을 최적화 할 때도 음소의 다양성을 고려해야 한다. 이를 위해 먼저 텍스트 코퍼스에 G2P 변환을 적용하여 음소 정보를 확보하였다. 일상 대화에서 자주 사용하는 발화에는 특정 음소 단위들이 타 음소 단위보다 상대적으로 많이 나타난다. 모든 음소가 균일하게 나타나지 않고, 특정 음소만 많이 나타나는 분포를 가진다. 따라서 상대적으로 희소한 음소 단위인 외래어 음소열에 대해 충분한 학습을 하기 위해서는 방대한 양의 코퍼스가 필요하다. 그러나 많은 양의 코퍼스를 형성하는 것은 막대한 시간 및 비용이 소요되기에 현실적이지 않다. 때문에 본 연구에서는 다양한 운율 정보가 균형적으로 내포된 대량 코퍼스의 트라이폰 분포와 유사한 분포를 가지는 한정된 크기의 코퍼스를 구축하였다. Fig. 6과 Fig. 7에서 음식 주문 상황 대화와 일반대화 코퍼스의 트라이폰 분포를 확인할 수 있다.

Table 1. Examples of general conversation and menu-ordering conversation.

General conversation	아니 다 처음 보는 사람들 이어서 그러는 거 아니야. 처음 그니까 그런 면접을 처음 보는 사람들이어서
Menu-ordering conversation	비프 스트로가노프 열 세개 주문을 받아주십시요

최적 크기 코퍼스 구축을 위한 코퍼스 추출 과정은 다음과 같다. 우선, AI hub에 존재하는 50만 문장의 일반 대화 코퍼스를 일반 대화 전체 코퍼스로 정의하고 해당 코퍼스의 트라이폰의 분포를 파악하였다. 이후 Balancer-Scripts[2]를 통해 유사한 트라이폰 분포를 가지는 1만 문장의 일반 대화 하위 코퍼스를 추출하였다. 전체 코퍼스와 하위 코퍼스의 분포 유사성은 피어슨 유사도 계산을 통해 검증하였다. 직접 구축한 14만 문장의 음식 주문 상황 대화 전체 코퍼스에도 앞과 동일한 방식을 적용하여 1만 문장의 음식 주문 상황 대화 하위 코퍼스를 추출하였다.

다음으로, 추출한 코퍼스의 트라이폰 분포 균형을 유지하면서 데이터셋 크기 최적화를 진행하였다. 먼저, 1,000문장에서 10,000문장까지 1,000문장 간격으로 음식 주문 상황 대화 하위 코퍼스 크기를 조절해 가며 일반 대화 하위 코퍼스와 병합하였다. 이어서, 병합한 코퍼스와 음식 주문 상황 대화 전체 코퍼스의 트라이폰 분포 유사성을 파악하기 위해 분포 간 피어슨 유사도를 계산하였다. 마지막으로, 가장 높은 피어슨 유사도를 가지는 경우의 음식 주문 상황 대화 코퍼스의 분량을 학습에 사용할 음식 주문 상황 대화 코퍼스 분량으로 선정하였다. 그 결과 일반 대화 1만 문장과 음식 주문 상황 대화 3,000문장을 병합한 13,000문장을 최종 코퍼스로 결정하였다. Table 1은 코퍼스 내 일반 대화 문장과 음식 주문 상황 대화 문장의 예시를 보여준다.

2) Balancer-Script: https://github.com/CShulby/Balancer-Scripts

4.2 모델 훈련

훈련 방법은 Espnet의 LJ speech 기준 훈련 방법을 따랐다.[3] 훈련 데이터는 앞서 4.1절에서 설명한 데이터셋을 사용하였다. 본 연구팀에서 자체 개발한 G2P 변환을 통해 음소열을 입력값으로 사용하였다. 음성-텍스트 쌍을 12500개, 250개, 250개로 나누어 훈련, 검증, 평가 데이터셋으로 사용하였다. 각 데이터셋에서 일반대화와 음식 주문 상황 대화의 비율을 전체 코퍼스의 비율인 10:3으로 유지하여 데이터 불균형에 따른 성능 저하를 방지하였다.

4.3 성능 평가

텍스트에서 음성을 합성하고 모델의 성능 평가를 위해 사용하는 컴퓨터 환경은 Ubuntu 20.04.3 LTS 64bit 운영체제, 64GB 메모리, Nvidia tesla v100 DGX-station 32GB GPU 1개이다.

4.4 MOS 평가 방법

평가자 20명이 40개 음성을 듣고 5점 만점으로 평가하였다. 일반대화 20개, 음식 주문 상황 대화 20개를 듣게 된다. 해당 음성은 모델 훈련과 검증에서 사용되지 않은 평가 데이터셋 250개에서 선정한 40개 음성이다. 평가자들은 주어진 원본 음성을 먼저 듣고 원본 음성과 비교하여 합성음의 발음 정확성을 평가한다.

4.5 DMOS 평가 결과

FastSpeech2와 콘포머 기반 FastSpeech2가 생성한 멜 스펙트로그램을 ParallelWave GAN[4]을 이용하여 음성으로 변환한 결과는 위와 같다(Table 2). 트랜스포머를 사용한 FastSpeech2에 비해 콘포머 기반 FastSpeech2가 월등한 성능을 보였다. FastSpeech2보다 콘포머 기반 FastSpeech2가 보다 정확한 발음을 구사하는 것을 확인할 수 있었다. 특히 콘포머 기반 FastSpeech2의 합성음과 원본 음성의 품질의 차이도 근소함을 확인할 수 있다.[4]

3) LJ speech training template from Espent : https://github.com/espnet/espnet/tree/master/egs2/TEMPLATE/tts1

4) audio samples available at : https://sogang-isds.github.io/korean-conformer-tts/

Table 2. DMOS scores of FastSpeech2 and conformer-based FastSpeech2.

	DMOS
Ground Truth	4.95 (±0.06)
FastSpeech2	3.72 (±0.18)
Conformer-based FastSpeech2	**4.04** (±0.50)

Table 3. Comparsion of DMOS scores between general conversation and menu-ordering convertsation.

DMOS	General conversation	menu-ordering conversation
Ground Truth	4.9335 (±0.10)	4.9662 (±0.16)
FastSpeech2	3.6675 (±0.18)	3.7745 (±0.49)
Conformer-based FastSpeech2	**4.0402** (±0.64)	**4.0450** (±0.95)

4.5.1 대화별 DMOS 평가 결과 비교

본 논문에서 제안한 합성기의 일반대화에서의 성능과 음식 주문 상황 대화의 성능을 비교한다. FastSpeech2의 경우 일반대화 20개를 평가했을 때와 음식 주문 상황 대화 20개를 평가했을 때[5]의 DMOS 점수의 차이가 0.107이다. 콘포머 기반 FastSpeech2는 두 대화의 점수 차이가 0.0048로 더 근소했다(Table 3).

본 논문에서는 고유어 기반의 일반대화와 외래어 기반의 음식 주문 상황 대화를 적절한 비율로 구성한 코퍼스를 제작하였다. 그리고 FastSpeech2와 콘포머 기반 FastSpeech2의 학습을 해당 코퍼스로 진행하였다. 그 결과 대화의 종류와 관계없이 일정한 성능을 낸 것을 확인할 수 있다. 특히 주로 외래어로 구성된 음식 주문 상황 대화에서도 성능이 저하되지 않았다. 일반대화와 음식 주문 상황 대화 모두 콘포머 기반의 FastSpeech2가 FastSpeech2에 비해 좋은 성능의 음성을 합성해냈음을 확인하였다.

V. 결 론

본 논문에서는 콘포머 기반 FastSpeech2를 이용한 한국어 음식 주문 문장 음성합성기를 제안하였다. 콘포머는 트랜스포머와 합성곱 신경망을 결합한 구

5) 4.4절에서 언급한 40개의 음성과 같은 음성이다.

조다. 트랜스포머에서와 달리 순방향 신경망을 반으로 나누어 전체 구조를 감싸게 하였고, 멀티 헤드 셀프 어텐션 모듈을 지난 후 합성곱 신경망을 배치하였다. 음성인식 분야에서 콘포머 구조를 이용하여 인식률 성능을 개선한 것이 확인되었다.

본 연구에서는 기존 음성인식 분야에서 좋은 성능을 낸 콘포머를 음성합성 분야에도 적용하였다. 이를 위하여 외래어 음소열을 포함하며 음소 균형적인 데이터셋을 제작하고, 해당 데이터를 활용하여 음성합성 모델을 훈련하였다. 음성합성 모델에서의 콘포머의 기여도를 실험하기 위하여 FastSpeech2와 콘포머 기반 FastSpeech2를 훈련하고, 성능을 비교하였다. 두 모델의 차이점은 FastSpeech2는 트랜스포머를, 콘포머 기반 FastSpeech2는 콘포머를 사용한 것이고 이 외의 구조는 동일하다. 실험 결과 트랜스포머 기반 FastSpeech2보다 콘포머 기반 FastSpeech2를 사용했을 때 좋은 품질의 합성음을 생성해낼 수 있음을 확인하였다. 추가적으로 일반대화와 음식 주문 상황 대화에 대한 합성음 음질을 비교하였다. 본 논문에서 제작한 코퍼스로 학습을 진행했을 때 외래어가 많은 음식 주문 문장에서도 합성음의 품질이 저하되지 않는 것을 확인하였다.

앞으로의 연구는 텍스트 분야에 콘포머 구조를 적용해 보는 방향으로 나아갈 수 있다. 음성 분야에서 콘포머를 사용하여 성능 향상을 확인하였고, 텍스트 분야에서도 성능 향상을 기대해 볼 수 있다.

또한 본 논문에서 제안한 음소열의 분포를 고려한 희소 음소열 특화 코퍼스 구축 방법을 활용하여 범용성을 가지면서도 다양한 특정 상황에 특화된 음성합성기 개발에 사용될 수 있을 것이다.

감사의 글

이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2022-0-00621,대화 기반 설명가능성을 멀티모달로 제공하는 인공지능 기술 개발).

References

1. A. J. Hunt and A. W. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” Proc. IEEE ICASSP, 373-376 (1996).
2. T. Yoshimura, K. Tokuda, T. Masuko, T, Kobayashi, and T. Kitamura, “Simultaneous modeling of spectrum, pitch and duration in HMM based speech synthesis,” Proc. Eurospeech, 2347-2350 (1999).
3. 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,” Proc. IEEE ICASSP, 4779-4783 (2018).
4. Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T-Y. Liu. “Fastspeech2: Fast and high-quality end-to-end text to speech,” arXiv:2006.04558 (2021).
5. A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “Wavenet: a generative model for raw audio,” arXiv:1609.03499 (2016).
6. R. Yamamoto, E. Song, and J. Kim. “Parallel wave-GAN: A fast waveformgeneration model based on generative adversarial networks with multi-resolution spectrogram,” Proc. IEEE ICASSP, 6199-6203 (2020).
7. Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu. “Fastspeech:Fast, robust and controllable text to speech,” Proc. NIPS, 3165-3174 (2019).
8. A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Yu Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, “Conformer: Convolution-augmented transformer for speech recognition,” Proc. Interspeech, 5036-5040 (2020).
9. M. Koo, “A korean speech recognition based on conformer” (In Korean), J. Acoust. Soc. Kr. **40**, 488-495 (2021)
10. P. Guo, F. Boyer, X. Chang, T. Hayashi, Y. Higuchi, H. Inaguma, N. Kamo, C. Li, D. Garcia-Romero, J. Shi, J. Shi, S. Watanabe, K. Wei, W. Zhang, and Y. Zhang, “Recent developments on espnet toolkit boosted by conformer,” Proc. IEEE ICASSP, 5874-5878 (2021)
11. N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” Proc. NeurIPS, 1-11 (2017)
12. P. Ramachandran, B. Zoph, and Q. V. Le, “Swish: A self-gated activation function,” arXiv:1710.05941v1 (2017).

저자 약력

▸ 최 예 린 (Yerin Choi)

2022년 2월 : 서강대학교 경제학과 학사, 컴퓨터공학과 학사
2022년 3월 ~ 현재 : 서강대학교 인공지능학과 대학원 석사과정

▸ 장 재 후 (JaeHoo Jang)

2022년 2월 : 건국대학교 산업공학과 학사
2022년 3월 ~ 현재 : 서강대학교 인공지능학과 대학원 석사과정

▸ 구 명 완 (Myoung-Wan Koo)

1982년 2월 : 연세대학교 전자공학과 학사
1985년 2월 : 한국과학기술원 전기및전자공학과 석사
1991년 2월 : 한국과학기술원 전기및전자공학과 박사
1985년 4월 ~ 2012년 7월 : KT 상무보
1996년 11월 ~ 1997년 12월 : 미국 벨연구소 방문연구원(연구재단 post-doc fellowship)
2012년 8월 ~ 현재 : 서강대학교 컴퓨터공학과 교수
2022년 3월 ~ 현재 : 서강대학교 인공지능학과 대학원 교수