pISSN 2005-8063
eISSN 2586-5854
2020. 9. 30.
Vol.12 No.3
pp. 55-63

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2020.12.3.055

Voice-to-voice conversion using transformer network*

June-Woo Kim · Ho-Young Jung**

Department of Artificial Intelligence, Kyungpook National University, Daegu, Korea

Abstract

Voice conversion can be applied to various voice processing applications. It can also play an important role in data augmentation for speech recognition. The conventional method uses the architecture of voice conversion with speech synthesis, with Mel filter bank as the main parameter. Mel filter bank is well-suited for quick computation of neural networks but cannot be converted into a high-quality waveform without the aid of a vocoder. Further, it is not effective in terms of obtaining data for speech recognition. In this paper, we focus on performing voice-to-voice conversion using only the raw spectrum. We propose a deep learning model based on the transformer network, which quickly learns the voice conversion properties using an attention mechanism between source and target spectral components. The experiments were performed on TIDIGITS data, a series of numbers spoken by an English speaker. The conversion voices were evaluated for naturalness and similarity using mean opinion score (MOS) obtained from 30 participants. Our final results yielded 3.52±0.22 for naturalness and 3.89±0.19 for similarity.

Keywords: voice conversion, transformer network, signal-to-signal conversion

1. 서론

최근 인공지능 기술이 발전함에 따라 많은 기업에서 음성 변환, 음성 합성 등의 서비스를 선보이고 있다. 음성 변환은 입력 목소리를 변환 목표 목소리로 음성의 내용을 유지한 채 변환하는 것을 의미한다. 음성 변환은 특정 음성 신호가 모델의 입력으로 주어지면, 단시간 푸리에 변환(short-time Fourier transform, STFT)을 통해 주파수/시간 도메인의 스펙트럼을 얻은 후 변환 모델을 통해 이루어지는 것이 효과적이다. 스펙트럼을 사용하는 이유는 스펙트럼이 변환을 위한 정보를 멜 필터뱅크(mel filterbank) 대비 더 많이 보유하고 있기 때문이다. 특히 음성합성 및 음성인식에서 사용되는 전통적인 멜 필터뱅크를 구하여 음성 변환에 적용할 수 있는데, 이것은 계산 및 모델 학습의 편의성을 제공하는 장점이 있지만, 위상 정보가 제거된 멜 필터뱅크를 바로 음성으로 변환할 수 없기 때문에 보코더를 필요로 한다. 보코더는 음성 변환 및 합성의 품질 개선에 효과적이지만, 변환 음성의 다양성을 확보하기에 어려움이 있다.

최근 많은 연구에서 순차적인 데이터 모델링 작업을 위해

* This work was supported by Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government (MSIT) (2016-0-00564, Development of Intelligent Interaction Technology Based on Context Awareness and Human Intention Understanding).

** hoyjung@knu.ac.kr, Corresponding author

Received 30 July 2020; Revised 17 September 2020; Accepted 17 September 2020

long short-term memory(LSTM; Hochreiter & Schmidhuber, 1997), bidirectional long-short term memory(Bi-LSTM; Schuster & Paliwal, 1997), gated recurrent unit(GRU; Chung et al., 2014)와 같은 recurrent neural network(RNN)를 이용하여, 인코더-디코더 구조를 가지는 sequence to sequence(Seq2Seq; Sutskever et al., 2014) 형태의 뉴럴 네트워크 모델을 사용하여 음성뿐만 아니라 자연어 처리 분야에서 우수한 연구 결과를 도출하고 있다. RNN은 기계 번역 및 언어 모델링과 같이 순차적인 데이터 모델링 작업에 널리 사용되는 방식이다.

그러나, RNN은 순차적으로 한 컨텐츠씩 처리하기 때문에 병렬화에 문제가 있고 학습 속도가 느릴 수 밖에 없다. 데이터의 길이가 길어질수록 모델은 멀리 있는 위치의 컨텐츠를 잊어버리거나 다음 위치의 컨텐츠와 혼동할 수 있는 경향이 있다.

이 문제를 해결하기 위해 transformer 네트워크(Vaswani et al., 2017) 가 자연어 데이터 학습을 위해 제안되었다. Transformer 네트워크는 입력 및 출력 간의 전구역 종속성을 도출하기 위해 주의집중 메커니즘에 전적으로 의존하는 뉴럴 네트워크 모델 아키텍처이다. 데이터의 길이가 길어지는 경우 성능이 저하되는 기존 RNN의 단점을 자기 주의집중(self-attention) 기법으로 해결하였다. Convolutional neural network(CNN; Kim, 2014) 및 RNN 구조를 포함하지 않는 transformer 네트워크 기반 모델은 RNN의 단점을 해결함과 동시에 기존 자연어 처리용 뉴럴 네트워크 모델 대비 적은 학습 시간으로 우수한 성능을 기대할 수 있는 장점을 보여주었다(Vaswani et al., 2017). 기계 번역에서 최고 성능을 보였고, transformer 네트워크의 출현으로 대규모 자연어 데이터에 대한 빠른 학습이 가능해졌다. 또한, 빅 데이터를 사전 학습한 bidirectional encoder representation from transformers(BERT; Devlin et al., 2018), generative pre-trained transformer(GPT; Radford et al., 2018) 등의 사전 학습된 transformer 네트워크 기반의 모델들이 등장하였다. 특히 BERT는 번역뿐만 아니라 문장 요약, 문장 간의 관련성 예측 등 많은 자연어 처리 분야에서 널리 사용되고 있다.

본 논문에서는 순차적 데이터의 변화에 효과적인 transformer 네트워크 모델을 활용하여 음성 스펙트럼 자체의 변환을 통해 음성 변환을 수행하는 종단간(end-to-end) 음성 변환 방법을 제안한다. 보코더를 통한 음성 합성 방식이 아닌 원형 스펙트럼 레벨에서의 변환을 진행하는 직접적인 음성대 음성 변환 뉴럴 네트워크 모델을 개발한다. 제안된 방법은 보코더를 사용하지 않음으로써 일정 규모의 변환쌍 데이터로 학습을 수행하면, 다양한 음성 데이터를 얻기 위해 여러 개의 보코더를 따로 개발하는 것보다 효과적으로 음성 데이터의 다양성을 확보할 수 있고, 이것은 실세계에서 여러 목적의 서비스에 활용될 수 있을 것이다.

본 논문에서는 TIDIGITS(Leonard & Doddington, 1993) 데이터를 활용하여 연속적인 숫자 음성 도메인에 대한 음성 변환 성능을 평가한다. 본 논문은 개별 숫자 단위로 음성 변환 모델을 학습하고, 변환을 수행하는 디코딩 과정에서는 입력 음성에 대해 강제 정렬을 적용하여 개별 숫자 구간을 구한 후 start-of-sentence(SOS) 토큰을 적용한다. SOS 토큰을 적용하여 입력 음성의 변환 디코딩 과정에서 SOS 토큰마다 디코더를 초기화하여 개별 숫자 단위의 음성 변환이 연속적으로 이루어지도록 구성되었다.

대부분의 음성 인식 시스템은 성인 남성과 여성 데이터를 기반으로 개발되고 있다. 따라서 성인 남녀가 아닌 노인, 어린이 및 말장애자의 음성을 인식하는데 좋은 성능을 기대하기 어렵다. 특히, 아동 혹은 노인 음성은 음향 및 언어적 상관성의 절댓값과 다양성 측면에서 성인 음성과는 상당히 다르며, 아동의 음향 공간은 넓고 중첩되는 음소 클래스를 보인다(Potamianos et al., 1997). 기존 연구에서는(Kwon et al., 2016) 20대 남녀의 음성 인식률이 94%인 것에 비하여 노인들의 음성 인식률은 76%인 것을 보여주고 있다. 또한, 노인 음성의 평균 말하는 속도는 20대 화자에 비하여 약 20% 이상의 속도가 느리다는 연구 결과가 제시되고 있다.

따라서, 어린이 및 노인 음성을 성인 음성으로 변환하는 기술의 개발은 성인 중심의 음성 인식 시스템을 어린이 및 노인 등이 활용할 수 있도록 하는데 중요한 역할을 할 수 있을 것이다.

2. 관련 연구

본 장에서는 음성 변환을 위한 재귀적 신경망 기반 방법과 transformer 네트워크 기반 방법에 대한 기존 기술을 소개한다.

2.1. 재귀적 신경망 기반 음성 변환 방법

2.1.1. Translatotron

본 방법은 서로 다른 언어 간의 음성 번역을 한 번에 진행하는 것이다(Jia et al., 2019) . 딥러닝 기반 음성 인식, 기계 번역, 음성 합성에 대해서는 여러 기술들이 개발되고 있지만, 이 과정을 한 번에 처리하는 연구는 처음이다. 심지어 화자 특징 벡터를 이용하여 음성 변환까지 가능하다.

Translatotron 방법은 입력 멜 필터뱅크를 처리하는 인코더 부분과 목표 멜 필터뱅크를 생성하는 주의집중 기반의 디코더인 Seq2Seq 구조로 구성되어 있다. 인코더의 입력은 80차원의 로그 멜 필터뱅크를 사용하고, 8계층의 Bi-LSTM을 사용한다. 그리고, multi head attention(MHA)을 통해 인코더-디코더의 scaled-dot product attention이 이루어진다.

인코더에는 신호-신호 간의 변환이 잘 안 되는 것을 해결하기 위해 부가적인 인식 계층을 인코더 출력에 연결하였다. 즉, 멀티태스크 학습 기법을 활용하여 입력 음성을 음소 단위로 인식할 수 있는 인코더 출력을 얻어 음성 변환을 위한 의미있는 인코더를 학습하게 된다.

Translatotron의 디코더는 음성합성에서 사용되는 Tacotron 2(Shen et al., 2018)의 구조와 유사하다. 디코더에서 1025차원의 멜 필터뱅크의 프레임을 예측하며, 매 디코딩 스텝별로 2개의 필터뱅크 프레임을 예측한다. 보코더에서는 Griffin-Lim을 기본적으로 사용하였지만(Griffin & Lim, 1984), mean opinion score (MOS) 평가에서 음성의 자연성을 평가할 때 후처리 시간이 더

소요되는 WaveRNN을 사용함으로써 더 나은 결과를 얻을 수 있음을 제시하였다.

2.1.2. Parrotron

말장애 화자의 음성을 변환하기 위해 제안된 방법으로 해당 음성을 일반인의 음성으로 변환하는 기술을 개발하였고, 이를 통해 말장애 화자의 음성 인식 성능 개선이 가능함을 보였다(Biadsy et al., 2019).

Parrotron의 인코더 입력은 16 kHz의 샘플링 된 음성 파형으로부터 125-7,600 Hz의 범위에서 80채널 멜 필터뱅크를 추출하여 사용하였다. 인코더 입력의 특징 표현을 얻기 위해 32의 크기인 커널을 갖는 2계층의 CNN을 통과한 뒤, ReLU와 batch normalization을 적용한다. 통과된 벡터들은 1×3의 필터 크기를 갖는 convolutional Bi-LSTM을 거치는데, 이를 통해 매 프레임마다 주파수 축을 통해서 값들이 모이는 효과를 볼 수 있다. 그 후 256차원의 크기를 갖는 3계층의 LSTM을 통과하고, 다시 512차원의 크기로 변경되는 구조로 이루어진다.

RNN 구조인 디코더는 한 번의 디코딩 과정에서 1개 프레임씩 출력 스펙트럼을 예측하도록 구성된다. 이렇게 출력된 디코더의 결과 스펙트럼은 오디오 신호를 합성하기 위해 Griffin-Lim 알고리즘을 사용하여 예측된 크기와 일치하는 위상 정보를 추정한 뒤 역 단시간 푸리에 변환(inverse STFT)를 통해 음성 파형으로 변환된다. 신호-신호 간의 변환이 잘 안 되는 것을 해결하기 위한 보조 음소 인식 모듈은 학습 과정에만 인코더 출력에 적용되고, 실제 변환 과정에서는 사용되지 않는다.

2.2. Transformer 모델 기반 음성 변환 방법

본 장에서는 본 논문에서 기본 구조로 채택한 transformer 네트워크 모델 기반 음성 합성 및 음성 변환 관련 선행 연구에 대해 소개한다.

2.2.1. Transformer 네트워크 기반 음성 합성

본 방법은 기존 transformer 네트워크에 Tacotron 2의 장점을 결합하는 것이다(Li et al., 2019) . 입력 텍스트를 먼저 음소로 변환하는데, 입력 텍스트가 “This is the waistline.”일 경우, “dh . ih. s / ih. z / dh . ax / w . ey. s . t - l . ay . n / punc.” 와 같은 음소 정보로 변환된다.

다음으로 Tacotron 2의 유사한 구조를 사용하여 텍스트 임베딩 결과를 얻는다. 기존 Tacotron 2 대비 선형 투영(linear projection)이 추가되었고, transformer 네트워크의 위치 임베딩(postional embedding)이 추가되어 상대적 위치 또는 절대적 위치에 대한 정보를 알 수 있도록 하였다. 음성 합성에서 입력 텍스트와 대상 멜 필터뱅크의 스케일이 다를 수 있는 문제를 위치 임베딩을 사용하여 해결하였다.

제안된 방법은 Tacotron 2와 달리 transformer 네트워크의 인코더-디코더 구조를 적용하였다. Location sensitive attention 및 RNN을 사용하지 않고 MHA로 대체함에 따라 병렬 컴퓨팅으로 학습 속도를 높일 수 있고, 순차적인 프레임 데이터 사이의 관계를 모델링 할 수 있어 전체 맥락을 고려할 수 있게 되었다.

후처리 부분에서는 Tacotron 2와 마찬가지로, 멜 필터뱅크와 정지 토큰을 예측하기 위해 두 개의 서로 다른 선형 투영법을 사용하였고, WaveNET 보코더를 사용하여 멜 필터뱅크를 음성 파형으로 출력하였다.

2.2.2. Voice transformer network

본 방법은 transformer 네트워크 기반의 첫 음성 변환 사례이다(Huang et al., 2019) . 대규모 음성 합성 데이터를 학습한 음성 합성 모델로부터 지식을 전달받는 사전 학습 기법을 활용하여 음성 변환을 진행한 연구이다. 즉, 인코더와 디코더 각각 별도의 사전 학습을 진행하여 음성 합성 모델 학습 과정의 파라미터를 공유하게 된다.

디코더는 기존 음성 합성 모델을 학습하기 위한 대규모 음성 합성 데이터를 사용하여 사전 학습되었고, 이를 통해 은닉 표현 벡터(hidden representation)로부터 고품질 음성을 생성할 수 있게 된다. 인코더는 디코더에서 처리할 수 있는 은닉 표현 벡터로 입력 음성을 인코딩하도록 사전 학습되는 것으로, 사전 학습된 디코더를 고정한 상태에서 오토인코더(auto encoder) 스타일로 인코더를 학습하게 된다. 즉, transformer 네트워크 기반 음성 합성모델 파라미터에서 transformer 기반 음성 변환 모델 파라미터로 지식을 전달하는 구조로 볼 수 있다. 본 방법은 사전 학습된 모델의 매개 변수로 초기화된 음성 변환 모델이 제한된 학습 데이터로도 고품질의 음성을 생성할 수 있음을 보여주었다. 변환된 음성의 평가에는 WORLD 보코더가 사용되었다(Morise et al., 2016).

2.2.3. Voice conversion with transformer network

사전 학습된 단일 화자의 음성 합성 모델을 활용하여 transformer 네트워크 모델의 컨텍스트 보존 메커니즘 및 모델 적응을 활용하는 일 대 일 음성 변환하는 방법이다(Liu et al., 2020).

Guided attention loss를 활용하여(Tanaka et al., 2019) transformer 네트워크 내부 인코더-디코더 간의 자기 주의집중 성능을 개선하는 방법을 제안하였다. 또한, 컨텍스트 보존 메커니즘을 활용하여 학습 절차를 안정화하였다. 인코더의 출력으로부터 입력 음성을 복원하기 위한 소스 디코더에 시드(seed) 입력으로부터 목표 음성을 예측하기 위한 목표 디코더를 추가하여 사용하였다. 이를 통해 소스 디코더는 입력 음성의 언어 정보를 보존하도록 하는 효과를 보이고, 목표 디코더는 입력 음성과 목표 음성의 공유 정보를 인코딩하는 인코더를 얻도록 수행된다. WaveNet 보코더를 사용하여 LSTM 계열의 음성 변환보다 약 2.7배의 빠른 효과와 더 높은 MOS 성능을 보여주었다.

2.3. 선행 연구 대비 제안한 방법의 차별성

선행 음성 변환의 방법은 모두 음성 파형을 변환하기 위해 추가적인 계산이 필요한 보코더를 활용하였다. 보코더는 고품질의 음성 변환을 위해 효과적이지만, 다양한 음성 변환 응용 및 음성 인식을 위한 데이터 증강 등에 활용하는데 다양성이 부족

한 문제가 존재한다.

본 논문에서는 음성 변환의 목적과 순차적 데이터의 병렬 학습 처리 목적을 위해 transformer 네트워크 구조를 이용하는 직접적인 음성 대 음성 변환 방법을 제안한다. 스펙트럼 기반 음성 변환을 수행하여, 보코더를 사용하지 않고 원형 스펙트럼 자체에서 직접적인 음성 대 음성 변환을 수행하는 transformer 네트워크 기반 음성 변환 모델을 제안한다.

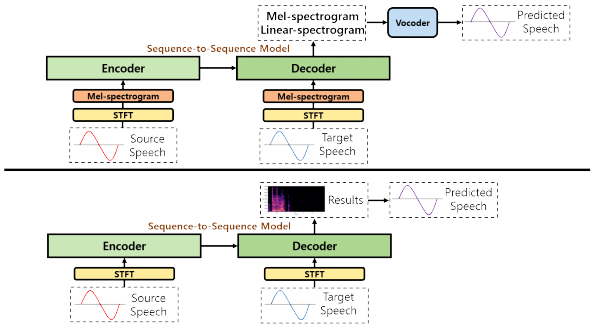

그림 1. Seq2Seq 모델 구조 기반의 음성 변환 및 합성의 선행 연구(상)와 본 논문에서 제안하는 모델 구조(하)

Figure 1. Conventional method of voice conversion and synthesis based on Seq2Seq model structure (Upper), our proposed model structure (down). Seq2Seq, sequence to sequence.

그림 1의 위쪽 부분을 보면 기존 음성 변환 및 합성 연구에서는 멜 필터뱅크를 이용하여 음성 변환을 진행하였는데, 스펙트럼보다 적은 파라미터로 인해 계산량을 줄일 수 있지만, 스펙트럼으로부터 압축될 때 신호 손실이 있기 때문에 음성 파형을 출력하기 위해서는 보코더가 필요하다.

본 연구에서는 그림 1의 아래쪽 부분과 같이 STFT를 이용하여 얻은 원형 스펙트럼을 모델의 입력으로 넣어 직접적인 음성 변환을 수행하고, 변환 모델의 출력 스펙트럼에 입력 화자의 위상을 적용한 후 inverse STFT를 통해 음성을 복원하는 방법을 사용한다.

3. Transformer 네트워크 기반 음성대 음성 변환

제안된 방법은 기존에 개발된 기술(Kim et al., 2020)에 기반하여 확장되었다. 기존 방법은 사전 학습된 모델 및 음성 합성의 개념을 사용하지 않고 transformer 네트워크 기반의 종단간 학습을 통하여 음성 변환이 가능함을 보여주었다. 개별 단어에 대하여 언어적 정보를 잃지 않은 채로 입력 화자와 대상 화자 간의 일 대 일 매핑을 통해 음성 변환이 가능함을 보여주었다. 이것은 다른 선행 연구 방법과 달리 보코더를 사용하지 않고도 음성 변환이 가능함을 보여주는 결과이다.

그러나, 변환 쌍별로 모델을 학습하는 단점을 가지고 있어, 어린이 음성을 성인 남성, 성인 여성의 음성으로 변환하려면 각각의 모델을 구축해야 하는 문제가 존재하게 된다. 따라서 본 논문에서는 음성 변환 transformer 모델에 화자 임베딩 정보를 추가하여 하나의 모델에서 여러 목표 음성으로 변환될 수 있는 universal 음성 변환 transformer 모델을 제안한다.

또한, 단어 단위로 학습된 변환 모델을 이용하여 연속발화의 자동 변환을 수행하기 위해 SOS 토큰을 기준으로 디코더를 초기화하여 연속적으로 이루어지는 구조로 개선하였다.

3.1. 연속 음성 변환

선행 학습된 강제 정렬기(McAuliffe et al., 2017)를 사용하여 연속적 단어로 이루어진 전체 음성에서 개별 단어의 경계를 추출한 후 각 경계마다 SOS 토큰을 추가하여 음성 변환 모델의 입력으로 적용한다. 이 경우 디코더는 SOS 토큰이 있을 때마다 초기화되어 한 단어 변환 후 다음 단어에 대해 새롭게 음성 변환을 수행하게 된다.

음성대 음성의 직접적인 변환을 학습하는 과정에서 동일 문장의 변환쌍 데이터를 일정 규모로 수집하는 것에 비해 개별 단어 단위의 변환쌍 데이터를 수집하는 것이 훨씬 수월하므로, 음성 변환의 학습 과정은 개별 단어 단위로 수행하는 것이 효과적이다. 따라서 여러 개의 단어로 이루어진 연속 음성을 변환하기 위해서는 개별 단어 단위의 디코딩이 요구되며, 제안된 방법은 SOS 토큰마다 디코더를 초기화하여 수행하는 구조로 이를 해결하게 된다

3.2. Universal 음성 변환 모델 구조

본 논문에서 제안한 모델 구조는 그림 2와 같다. 성인 남자 음성으로부터 성인 여자, 남자 어린이 그리고 여자 어린이의 음성으로 변환하는 과정을 예로 들면, 인코더의 입력은 남자의 음성인 x가 사용되고, 디코더의 입력은 같은 언어적 정보를 가지고 있는 목표 음성인 y를 입력으로 받아 학습이 진행된다. 각 입력단에서 음성 파형에 STFT를 취하여 스펙트럼을 얻고, 크기와 위상 정보로 분리한다. Transformer 네트워크에는 RNN 또는 CNN이 포함되어 있지 않으므로 위치 임베딩을 활용하여 입력 스펙트럼의 매 시간마다 상대적 위치에 대한 정보를 모델에 제공한다.

$$PE(pos,2i) = \sin(pos/10{,}000^{2i/d_{model}}) \quad (1)$$

$$PE(pos,2i+1) = \cos(pos/10{,}000^{2i/d_{model}}) \quad (2)$$

식 (1)과 (2)에서 PE는 위치 임베딩, pos는 입력 스펙트럼 벡터의 위치를 나타낸다. i는 스펙트럼 내의 차원의 인덱스를 의미하며, 주기가 $10{,}000^{2i/d_{model}} * 2\pi$ 인 삼각 함수이다. 본 논문에서는 d_{model}을 256차원으로 설정하여 실험을 진행하였기 때문에, 해당 스펙트럼은 i에 0부터 $d_{model}/2$인 128까지를 대입하는 과정을 통해 256차원의 위치 벡터를 얻게 된다. 위치 임베딩의 time step이 $2i$일 때는 사인 함수를 사용하고 $2i+1$일 때는 코사인 함수를 사용하게 된다.

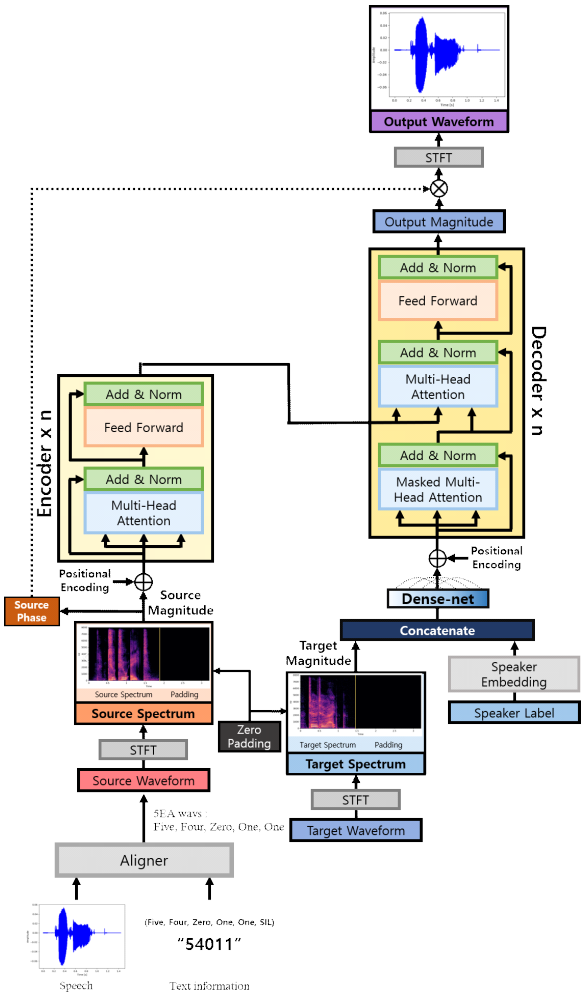

그림 2. 본 논문에서 실험에 사용한 시스템 아키텍처
Figure 2. Our proposed voice conversion model architecture

인코더의 경우 입력 스펙트럼 성분에 위치 임베딩을 적용하여 위치 벡터가 더해진 값이 최종 인코더 입력이 된다. 다음으로 transformer 네트워크의 MHA을 수행하며, ReLU가 포함된 2계층의 피드포워드 네트워크(feed-forward network)를 사용하여 벡터값들을 정규화한다. 이 과정을 통해 매 프레임별로 이루어진 순차적 정보 전체에 대한 새로운 컨텍스트 정보를 만들 수 있다. 이 값과 스펙트럼 크기 성분에 위치 임베딩이 더해진 입력값 사이에 요소별 합 연산(element-wise addition)을 진행하는 residual connection이 적용된다. 이에 대한 식은 아래와 같다.

$$y = h(x_l) + F(x_l, W_l) \tag{3}$$

$$x_{l+1} = f(y_l) \tag{4}$$

식 (3)과 (4)에서 x_l과 x_{l+1}은 입력과 출력의 l번째 단계이며, F는 residual 함수이고 f는 ReLU 함수이다. 이를 통해 컨텍스트 정보를 입력 데이터로부터 추출하여 활용할 수 있다. 이 과정을 통해 인코더는 주어진 입력 데이터 전체의 전역 종속성을 도출할 수 있으며 주의집중 메커니즘인 MHA을 통해 각 프레임 정보를 효과적으로 나타낼 수 있게 된다.

디코더에서는 인코더의 방식과 마찬가지로 대상 y의 파형에서 STFT를 적용한 스펙트럼 크기 성분을 사용하며, 다중 화자의 목소리로 음성 변환을 진행하기 위하여 원 핫 인코딩을 사용하여 화자 임베딩이 되도록 하였다. 32차원의 Dense Net에 softsign 함수를 활용하여 화자 임베딩 벡터 값을 만든 뒤, 목표 화자의 음성 y와 결합하게 된다. 이 경우 MHA의 은닉 노드 크기와 맞지 않기 때문에 Dense Net으로 차원을 일치하도록 하고, 여기에 위치 임베딩을 추가 적용하여 구성한다.

디코더는 인코더와 구조는 거의 비슷하지만, MHA에서 자기 주의집중을 수행할 때 mask를 적용한 점이 다르다. Masked MHA을 사용하는 이유는 자기 주의집중이 진행될 때, 현재 프레임 이후의 정보를 모델에서 활용하지 못하게 하는 것을 의미한다. 다음으로 인코더-디코더 간의 주의집중이 수행되는데, 디코더의 i번째 입력 y_i의 정보를 표현하기 위하여 인코더 입력 x의 정보를 이용하는 구조를 학습하게 된다. 최종적으로 디코더의 masked MHA 결과에 인코더-디코더 주의집중의 결과가 더해져서 피드포워드 네트워크의 입력이 되고 정규화 과정을 통해 최종 출력을 얻을 수 있다. 이 과정은 인코더, 디코더 구조의 적층 수 만큼 반복이 진행된다.

인코더와 디코더의 입력, 그리고 출력 $\hat{y}$의 d_{model}의 크기는 모두 같으며, $\hat{y}$에는 현재 변환된 대상의 스펙트럼 크기 정보만 포함하고 있다. 이를 변환된 음성 파형으로 만들어주기 위해 입력 음성의 위상을 적용한 complex 스펙트럼을 추정하여 inverse STFT를 통해 변환된 음성 파형을 얻을 수 있다.

학습에는 adam optimizer(Kingma & Ba, 2015)가 사용되었으며, $\beta_1 = 0.9$, $\beta_2 = 0.98$ 및 $\epsilon = 1e-9$의 값을 갖는다. 학습률은 $1e-4$로 시작하여 4,000 스텝마다 0.96%만큼 감소하도록 설정하였다. 또한, 6계층의 인코더와 디코더를 사용하였고, MHA의 head는 8개를 사용하였다. 인코더와 디코더에서 사용된 d_{model}의 크기는 256이고, 피드포워드 네트워크의 d_{model} 크기는 1,024이다. Dropout은 0.1로 설정하였고, 이는 학습 단계에서만 사용되었다. 모델의 학습은 L1 loss를 사용하여 진행되었다.

3.3. 학습 단계

학습 과정은 개별 단어 단위로 이루어진다. 'zero, one, ..., oh'까지의 총 11개의 단어에 대하여 transformer 네트워크를 사용하여 언어적 정보를 유지한 채 대상의 목소리 변환을 위한 학습을 진행하였다. 숫자 음성 단위별로 길이가 다르므로 전체 학습 데이터에서 최대 길이를 구하여 transformer 구조를 결정하고 짧은 데이터에 대해서는 제로 패딩(zero padding)을 적용하였다. 성인 남자 음성으로부터 성인 여자, 남자 어린이 및 여자 어린이 음성으로의 변환 과정을 예로 들면, 인코더의 입력은 성인 남자의 음성이 적용되고 디코더의 입력에는 같은 숫자를 발성한 성인 여자, 남자 어린이 및 여자 어린이의 음성에 원 핫 인코딩의 목표 화자 임베딩 값이 더해져서 적용된다.

디코더 입력의 가장 앞에 n개의 차원으로 이루어진 벡터값을 SOS 토큰으로 활용하여 변환 스펙트럼을 얻는 디코딩의 시작으로 활용하였다. 실제적으로는 본 연구의 입력 d_{model} 크기가 256이므로, (256, 1)의 크기를 갖는 0과 1 사이의 무작위 균등분포 값을 인위적으로 만들어 사용하게 된다. 또한, 교사 강제 학습(teacher-forcing learning)을 진행하기 위해 목표 데이터에 제로 패딩을 적용하는 시작점에 end of sentence(EOS) 토큰을 추가하여 모델 학습을 진행하였다. 제로 패딩 구간에 대하여 MHA의 자기 주의집중이 일어나지 않도록 $-1e^9$를 곱한 masked MHA을 적용하였다.

3.4. 추론 단계

추론 단계에서는 입력 음성에 대하여 언어적 정보를 잃지 않고 화자의 목소리만 변환하는 것이 목적이다. 그림 3과 같이 '54011'을 발성한 입력 음성 데이터와 이것의 단어 구간 정보를 이용하여 경계마다 SOS 토큰을 추가한 입력 스펙트럼 데이터를 구성하여 인코더의 입력으로 사용하게 된다.

디코더 입력으로는 시작을 알리는 SOS 토큰과 변환을 원하는 목표 화자의 원 핫 인코딩 기반 화자 임베딩 값이 주어진다. 화자 임베딩을 적용한 후 SOS 토큰과 같이 결합되어 음성 변환을 위한 transformer 디코딩이 시작된다.

연속적인 음성에서 개별 단어의 시작을 표시하는 SOS 토큰은 모델에게 새로운 음성 변환의 시작을 알리게 되며, 이를 통해 개별 단어 구간의 변환을 연속적으로 수행하는 음성대 음성 변환 결과를 얻을 수 있게 된다.

4. 실험 결과 및 분석

4.1. 음성 데이터 및 전처리

본 논문은 326명의 화자가 영문 숫자를 발성한 TIDIGITS 데이터를 사용하였다. 111명의 남자, 114명의 여자, 50명의 남자 어린이, 51명의 여자 어린이로 구성되어 있다. 동일 단어쌍을 이루는 55명의 남자, 57명의 여자, 25명의 남자 어린이, 26명의 여자 어린이의 학습 데이터를 사용하여 모델 학습을 진행하였다.

TIDIGITS 데이터는 20 kHz로 샘플링되었으며, 전용 녹음실 공간에서 전문 마이크로 녹음되었다. 본 논문에서는 20 kHz의 음성을 16 kHz로 다운 샘플링하여 진행하였다. 이 데이터로부터 얻은 스펙트럼 기반 입력은 (257, T)의 크기를 가지며, 여기서 T는 음성 프레임의 길이이다. 계산의 편의를 위해 마지막 스펙트럼 값을 제외하여 d_{model}는 256이 되었고, 최종적으로 (256, T)의 차원의 데이터를 구성하도록 전처리를 진행하였다.

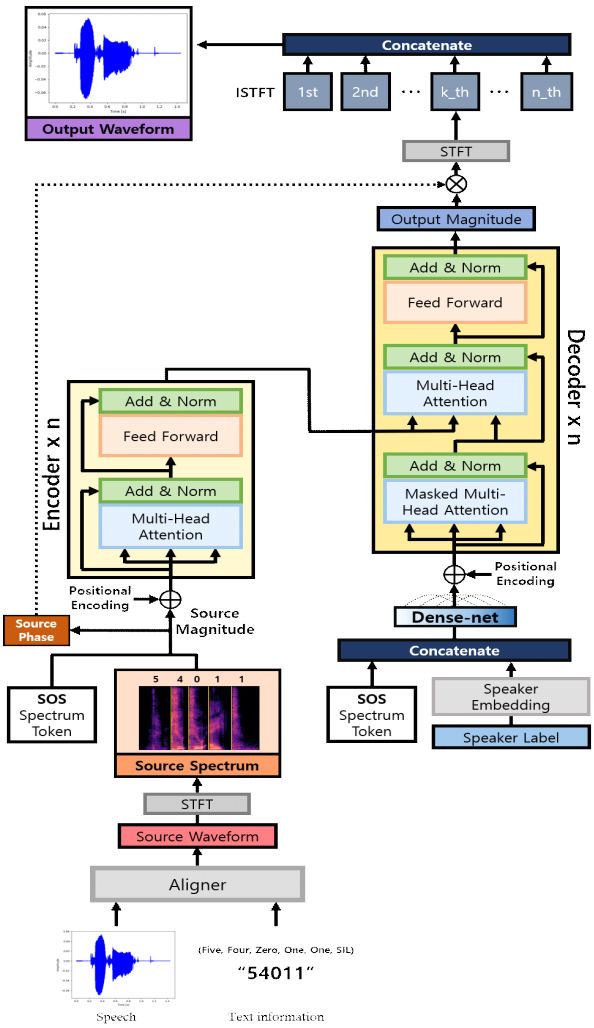

그림 3. 제안한 모델의 추론 단계 흐름
Figure 3. Our proposed model in evaluation task

4.2. 결과 및 분석

그림 4는 본 논문에서 제안한 음성 변환 방법의 결과이다. 입력 음성은 "228"을 발성한 것이다. 그림 4 내의 첫 번째 행은 음성 발화의 파형 형태이며, 두 번째 행은 STFT를 통한 스펙트로그램을 보여준다. 각 행별로 왼쪽은 모델의 입력인 여자 어린이의 음성이고, 중간은 변환 목표인 성인 여자의 음성이며, 오른쪽은 여자 어린이 음성에서 성인 여자 음성으로 변환된 결과이다.

그림 4의 파형과 스펙트로그램에서 볼 수 있듯이, 입력 여자 어린이의 음성은 에너지가 성인 여자보다 비교적 낮았고, 음성 발화의 시작 지점이 성인 여자의 음성 발화보다 약 0.2초 가까이 느린 특성을 보인다. 변환된 결과는 여자 어린이 음성을 성인 여자 음성으로 변환하면 성인 여자 음성의 에너지와 비슷해짐을 확인할 수 있고, 여자 어린이의 '2'와 '8' 사이의 연속적으로 겹쳐진 부분이 구별되어 변환된 모습을 확인할 수 있다.

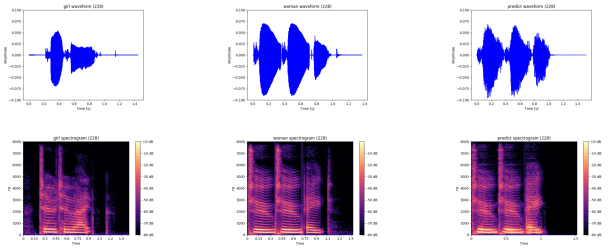

그림 4. 왼쪽은 입력인 여자 어린이의 발화, 중간 그림은 대상인 성인 여자의 발화, 마지막 오른쪽은 변환된 음성의 발화 결과이며 모두 228에 대해 영어('two-two-eight')로 발화한 결과이다.
Figure 4. Visualization of our voice conversion results. The left part of the figure shows the real input from a girl's voice. Middle part of the figure shows the target from a woman's voice and the right part of the figure shows the inference results of conversion from a girl's voice to a woman's voice saying "228" in English

제안된 방법의 평가를 위해 20대에서 30대 사이의 30명을 대상으로 변환된 목소리의 자연성 및 유사성에 대하여 MOS 평가를 진행하였다. MOS 평가는 자연성 및 유사성 1−5점 척도로 진행하였으며, 여기서 1점은 변환된 음성의 품질이 가능 낮은 경우이고 5점은 가장 높은 품질을 가지는 경우이다. 평가는 성인 남자, 성인 여자, 남자 어린이와 여자 어린이에 대해 무작위로 추출한 음성 변환 결과를 대상으로 진행하였다.

표 1. 음성 변환 자연성 MOS 평가 결과
Table 1. MOS naturalness evaluation result of our voice conversion

입력 음성 화자	대상 음성 화자			
	성인 남자	성인 여자	남자 어린이	여자 어린이
성인 남자	-	3.17±0.32	3.75±0.26	4.00±0.27
성인 여자	2.76±0.30	-	3.21±0.35	3.38±0.35
남자 어린이	3.03±0.31	3.10±0.35	-	4.10±0.22
여자 어린이	2.93±0.33	3.31±0.36	3.48±0.31	-
자연성 전체 평균	3.52±0.22			

MOS, mean opinion score.

표 1은 본 논문에서 제안된 음성 변환 방법의 자연성 평가 점수를 나타내고, 표 2는 유사성 평가 점수를 나타내고 있다. 첫 번째 집단인 성인 남자 음성의 경우는 입력 데이터가 성인 남자 음성이고 성인 여자, 남자 어린이, 여자 어린이의 음성이 목표인 경우에 대해 변환 성능을 평가한 것이다. 음성의 자연성과 유사성 평가 결과는 성인 남자 음성을 변환하는 경우 가장 좋은 성능을 보였다. 반면에, 성인 여자 음성을 변환한 결과가 다른 경우에 비해 낮은 성능을 보임을 알 수 있다. 본 논문에서 제안된 음성 변환 모델은 자연성 3.52±0.22 및 유사성 3.89±0.19의 성능으로, transformer 기반 음성대 음성 변환이 가능함을 보여주었다.

표 2. 음성 변환 유사성 MOS 평가 결과
Table 2. MOS similarity evaluation result of our voice conversion

입력 음성 화자	대상 음성 화자			
	성인 남자	성인 여자	남자 어린이	여자 어린이
성인 남자	-	3.72±0.29	4.17±0.22	4.48±0.21
성인 여자	3.28±0.29	-	3.66±0.32	3.97±0.31
남자 어린이	3.28±0.30	3.59±0.28	-	4.69±0.17
여자 어린이	3.48±0.33	4.14±0.27	4.07±0.25	-
유사성 전체 평균	3.89±0.19			

MOS, mean opinion score.

5. 결론

멜 필터뱅크는 원형 스펙트럼보다 비교적 낮은 차원의 주파수로 구성이 되어있다. 따라서, 뉴럴 네트워크 학습의 편리성 및 빠른 연산 속도를 제공하지만 보코더 없이 음성 파형으로 변환할 수 없는 문제가 있다. 보코더를 사용하면 음성 변환 응용이나 음성 인식을 위한 데이터 증강 등의 실질적인 문제를 해결하는데 필수적인 다양성의 확보에 제한이 있다. 이런 문제를 해결하기 위해 본 논문은 STFT를 통과한 원형 스펙트럼 자체에 초점을 두어 직접적인 음성대 음성의 변환을 수행하는 뉴럴 네트워크 모델을 제안하였다. 또한, 변환 쌍별로 각각의 모델을 가져야 하는 문제를 해결하기 위해 음성 변환 모델에 화자 임베딩 정보를 추가하여 하나의 모델에서 여러 목표 음성으로 변환할 수 있는 universal 음성 변환 transformer 모델을 제안하였고, 연속적인 단어 발성에 대해 자동 변환이 가능하도록 디코딩 구조를 개선하였다.

본 논문은 30명의 평가자를 모집하여 제안된 방법의 성능을 자연성과 유사성에 대해 평가하였다. 제안된 방법은 자연성 3.52±0.22, 유사성 3.89±0.19의 성능으로 음성대 음성의 직접적인 변환이 가능함을 제시하였다.

References

Biadsy, F., Weiss, R. J., Moreno, P. J., Kanvesky, D., & Jia, Y. (2019). Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation. *arXiv.* Retrieved from: https://arxiv.org/abs/1904.04169

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv.* Retrieved from: https://arxiv.org/abs/1412.3555

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv.* Retrieved from: https://arxiv.org/abs/1810.04805

Griffin, D., & Lim, J. (1984). Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *32*(2), 236-243.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory.

Neural Computation, *9*(8), 1735-1780.

Huang, W. C., Hayashi, T., Wu, Y. C., Kameoka, H., & Toda, T. (2019). Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining. *arXiv.* Retrieved from: https://arxiv.org/abs/1912.06813

Jia, Y., Weiss, R. J., Biadsy, F., Macherey, W., Johnson, M., Chen, Z., & Wu, Y. (2019). Direct speech-to-speech translation with a sequence-to-sequence model. *arXiv.* Retrieved from: https://arxiv.org/abs/1904.06037

Kim, J. W., Jung, H. Y., & Lee, M. (2020). Vocoder-free end-to-end voice conversion with transformer Network. *arXiv.* Retrieved from: https://arxiv.org/abs/2002.03808

Kim, Y. (2014, October). Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1746-1751). Doha, Qatar.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *arXiv.* Retrieved from: https://arxiv.org/ abs/1412.6980

Kwon, S., Kim, S. J., & Choeh, J. Y. (2016). Preprocessing for elderly speech recognition of smart devices. *Computer Speech & Language*, *36*, 110-121.

Lee, J., Cho, K., & Hofmann, T. (2017). Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, *5*, 365-378.

Leonard, R. G., & Doddington, G. R. (1993). *Tidigits speech corpus.* Philadelphia, PA: Texas Instruments.

Li, N., Liu, S., Liu, Y., Zhao, S., & Liu, M. (2019, July). Neural speech synthesis with transformer network. *Proceedings of the 33rd AAAI Conference on Artificial Intelligence* (Vol. 33, pp. 6706-6713). Hawaii, HI.

Liu, R., Chen, X., & Wen, X. (2020, May). Voice conversion with transformer Network. *Proceedings of the ICASSP 2020－2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7759-7759). Barcelona, Spain.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017, August). Montreal forced aligner: Trainable text-speech alignment using Kaldi. In *Interspeech* (Vol. 2017, pp. 498-502). Stockholm, Sweden.

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.

Potamianos, A., Narayanan, S., & Lee, S. (1997, September). Automatic speech recognition for children. *Proceedings of the 5th European Conference on Speech Communication and Technology* (pp. 2371-2374). Rhodes, Greece.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. Retrieved from https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper.pdf

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673-2681.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., ... Saurous, R. A. (2018, April). Natural tts synthesis by conditioning wavenet on MEL spectrogram predictions. *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4779-4783). Calgary, AB.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems 27 (NIPS 2014)* (pp. 3104-3112). San Mateo, CA.

Tanaka, K., Kameoka, H., Kaneko, T., & Hojo, N. (2019, May). AttS2S-VC: Sequence-to-sequence voice conversion with attention and context preservation mechanisms. *Proceedings of the ICASSP 2019－2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6805-6809). Brighton, UK.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008). San Mateo, CA.

• **김준우 (June-Woo Kim)**
경북대학교 인공지능학과 석사과정
대구광역시 북구 대학로 80 경북대학교
Tel: 053-940-8616
Email: kaen2891@gmail.com
관심분야: 음성 인식, 음성 변환, 음성 합성, 딥러닝

• **정호영 (Ho-Young Jung)** 교신저자
경북대학교 인공지능학과 교수
대구광역시 북구 대학로 80 경북대학교
Tel: 053-950-2337
Email: hoyjung@knu.ac.kr
관심분야: 음성 인식, 음성 변환, 자연어 이해, 딥러닝

Transformer 네트워크를 이용한 음성신호 변환*

김 준 우 · 정 호 영

경북대학교 인공지능학과

국문초록

음성 변환은 다양한 음성 처리 응용에 적용될 수 있으며, 음성 인식을 위한 학습 데이터 증강에도 중요한 역할을 할 수 있다. 기존의 방법은 음성 합성을 이용하여 음성 변환을 수행하는 구조를 사용하여 멜 필터뱅크가 중요한 파라미터로 활용된다. 멜 필터뱅크는 뉴럴 네트워크 학습의 편리성 및 빠른 연산 속도를 제공하지만, 자연스러운 음성 파형을 생성하기 위해서는 보코더를 필요로 한다. 또한, 이 방법은 음성 인식을 위한 다양한 데이터를 얻는데 효과적이지 않다. 이 문제를 해결하기 위해 본 논문은 원형 스펙트럼을 사용하여 음성 신호 자체의 변환을 시도하였고, 어텐션 메커니즘으로 스펙트럼 성분 사이의 관계를 효율적으로 찾아내어 변환을 위한 자질을 학습할 수 있는 transformer 네트워크 기반 딥러닝 구조를 제안하였다. 영어 숫자로 구성된 TIDIGITS 데이터를 사용하여 개별 숫자 변환 모델을 학습하였고, 연속 숫자 음성 변환 디코더를 통한 결과를 평가하였다. 30명의 청취 평가자를 모집하여 변환된 음성의 자연성과 유사성에 대해 평가를 진행하였고, 자연성 3.52±0.22 및 유사성 3.89±0.19 품질의 성능을 얻었다.

핵심어: 음성 변환, 트랜스포머 네트워크, 신호 대 신호 변환

* 본 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (2016-0-00564, 사용자의 의도와 맥락을 이해하는 지능형 인터랙션 기술 연구개발).