콘텐츠 내 텍스트 문맥 정보 활용을 위한 Dual-Path Cross-Attention 기반 텍스트-오디오 음성 분리 기술

(Context-Aware Dual-Path Cross-Attention for Content-Oriented Text-Audio Speech Source Separation)

이건우*
(Geon Woo Lee)

요 약

본 논문에서는 텍스트 데이터의 문맥 정보 반영을 위해 오디오 및 텍스트 데이터를 활용한 트랜스포머 기반 음성 분리 모델을 제안한다. 음성 분리 기술은 음성 신호와 배경 음원이 혼합된 미디어 콘텐츠 오디오 신호에서 음성 신호만을 분리할 수 있으며, 이와 같이 분리된 음성 신호는 콘텐츠 재가공을 위해 활용될 수 있다. 미디어 콘텐츠에 포함된 대본과 같은 텍스트 데이터는 문맥 정보를 잠재적으로 포함하여 음성 분리 모델의 성능을 향상하는 데 사용될 수 있다. 미디어 콘텐츠에 포함된 텍스트 데이터를 활용하기 위해 제안된 음성 분리 모델은 텍스트 임베딩 모델을 활용하여 텍스트 데이터에서 문맥 정보와 토큰 정보를 추출하고, 트랜스포머 기반 음성 분리 모델에 cross-attention을 적용하여 텍스트 임베딩 벡터와 오디오 특징 벡터 사이 정렬 정보가 학습되도록 구성한다. 제안된 음성 분리 모델은 LibriSpeech 및 MUSDB18 데이터셋을 기반한 시뮬레이션 데이터로 평가를 진행하였으며, 기존 텍스트 및 오디오 데이터를 사용하는 음성 분리 모델 대비 SDR, SAR, SIR, PESQ, STOI 지표에서 모두 우수한 성능을 나타냈다. 특히, 보컬 음원이 포함된 배경음악 환경에서도 텍스트 데이터를 활용함으로써 기존 방법보다 높은 음성 분리 성능을 달성했다. 이와 같은 텍스트-오디오 멀티 모달 기술은 미디어 콘텐츠 리믹싱 및 재생성 분야에서 주요한 역할을 수행할 것으로 기대한다.

■ 중심어 : 음성 분리 ; 텍스트-오디오 멀티 모달 ; Cross-attention ; Transformer 구조

Abstract

This paper proposes a transformer-based speech source separation model that leverages both audio and textual data to improve separation performance by incorporating contextual information from text. Speech separation enables the extraction of speech signals from audio mixtures containing background sources, facilitating downstream content repurposing. Textual information embedded in media content, such as scripts, provides contextual cues that can improve separation performance. To integrate such contextual cues, the proposed model employs a text embedding network to extract contextual and token-level representations, and integrates them into a transformer-based separation framework via cross-attention, enabling alignment between text and audio features. These features are then aggregated with the audio feature vector via cross-attention within a transformer-based separation model, which enables alignment between the two modalities. The proposed model is evaluated on simulated mixtures created from the LibriSpeech and MUSDB18 datasets. Experimental results demonstrate that the proposed model achieves performance improvements over existing text-audio separation models, as measured by SDR, SAR, SIR, PESQ, and STOI metrics. Furthermore, the proposed separation model achieves superior performance in scenarios involving vocal source-included background music, where conventional approaches typically degrade.

■ keywords : Speech source separation ; text-audio multi-modal ; Cross-attention ; Transformer architecture

Ⅰ. 서 론

최근 급속도로 발전하고 있는 인공지능 기술은 자연어, 비전뿐만 아니라 오디오 신호 처리 분야에서

* 정회원, 조선대학교 AI소프트웨어학부
이 논문은 조선대학교 학술연구비의 지원을 받아 연구되었음(2024년)
접수일자 : 2025년 07월 21일
수정일자 : 1차 2025년 08월 20일, 2차 2025년 08월 20일
게재확정일 : 2025년 08월 30일
단독저자 : 이건우 e-mail : geonwoo@chosun.ac.kr

도 활발하게 활용되고 있다. 오디오 신호 처리 분야 중 하나인 오디오 분리 기술은 음성 및 다양한 음향 신호가 혼합된 오디오 신호에서 각 신호를 분리하는 기술에 속하며, 음향 잡음 제거, 음성 분리 등 다양한 응용 분야에서 활용될 수 있다. 이와 같은 다양한 분야 중에서 음성 분리 기술은 주어진 오디오 신호에서 특정 음성 신호만을 분리하는 기술이다[1,2]. 음성 분리 기술은 일반적으로 잡음 또는 배경 음향 신호가 존재하는 환경에서 음성 신호 취득을 통해 음성 인식 성능 향상을 위해 사용될 수 있다[3,4]. 최근에는 개인 콘텐츠 창작의 시대가 열리면서 콘텐츠 리믹싱(remixing) 및 편집에 대한 수요가 증가하면서, 오디오 신호에서 특정 음성 분리가 요구사항이 발생할 수 있다[5]. 음성 분리 기술을 통해 배경 음향이 잘 제거된 음성 신호는 리믹싱 과정에서 높은 품질의 콘텐츠를 재생성할 수 있다[6].

음성 분리 기술은 다중 또는 단일 채널 오디오 신호를 대상으로 분리를 진행하며, 공간 정보를 상대적으로 많이 포함하고 있는 다중 채널 음성 분리 기술은 단일 채널 음성 분리 기술보다 상대적으로 좋은 성능을 달성할 수 있다[7]. 하지만, 대부분의 콘텐츠 플랫폼에서는 단일 또는 스테레오 채널로 오디오를 제공하고 있으므로 높은 품질의 분리된 음성 신호를 제공하기 어렵다. 또한, 콘텐츠 플랫폼에서 제공되는 스테레오 채널의 오디오는 단일 채널에서 업믹싱(upmixing) 되는 경우도 존재하여[8]. 실질적으로는 공간 정보를 다수 포함하지 못할 수 있다. 즉, 스테레오 채널을 사용하지만 실질적으로는 단일 채널을 사용함으로써 성능 개선에 한계가 존재할 수 있다. 즉, 일반적인 콘텐츠 플랫폼 환경에서 음성 분리 기술 적용을 위해서는 단일 채널 기반의 음성 분리 기술이 요구된다.

단일 채널 음성 분리 기술은 주어진 오디오 신호를 음향 신호와 음성 신호로 분리하는 기술이며, 이는 하나의 방정식에서 다수의 미지수를 추정하는 해가 유일하지 않은 문제로써 일반적인 수학적 접근 방법에는 한계가 존재한다. 이러한 특성으로 인해 전통적인 신호 처리 기법이나 선형 모델 기반의 기계 학습 접근 방법[9] 보다 비선형 관계 모델링에 탁월한 성능을 보이는 심층 학습 기법을 활용한 음성 분리 기술이 우수한 성능을 나타내고 있다[10-12].

일반적으로 널리 사용되는 합성곱 신경망 또는 재귀 신경망 기반 음성 분리 기술은 지역 정보를 활용하여 좋은 성능을 달성했다. 하지만, 오디오 신호의 길이가 긴 경우에는 오디오 신호의 전체 문맥 정보를 반영하는데 한계가 존재하여 상대적으로 낮은 성능을 나타낼 수 있다. 이와 같은 문제를 해결하기 위해 지역 정보뿐만 아니라 전역 정보를 활용할 수 있는 트랜스포머 기반 음성 분리 기술이 제안되었다. 더 나아가 SepFormer는 트랜스포머 기반 음성 분리 모델에서 시간 축과 주파수 축을 분리하여 분석하는 구조를 추가하여 일반적인 트랜스포머 기반 음성 분리 기술보다 좋은 성능을 나타냈다 [12].

일반적으로 미디어 콘텐츠 데이터는 단일 채널 오디오 신호만 활용할 수 있는 제약 환경이지만, 음성 신호 분리 기술에 활용할 수 있는 메타 데이터를 포함하는 경우가 존재한다. 예를 들어 영상 콘텐츠에는 자막, 해설 스크립트, 대본 등과 같은 텍스트 데이터가 포함된 경우가 존재하며, 이러한 텍스트 데이터는 오디오 신호의 특정 시점과 문맥의 내용을 함축적으로 포함할 수 있다. 최근에는 이러한 텍스트 데이터를 활용하여 음성 분리 성능 향상을 위한 다양한 연구가 진행되고 있다. 즉, 단순히 오디오 신호만을 활용하는 음성 분리 기술에 주어진 텍스트 데이터를 조건으로 활용함으로써 음성 분리 성능을 향상하는 다양한 연구가 진행되고 있다. 특히, 가사가 주어진 오디오 콘텐츠 등의 환경에서 효과적인 결과를 보였다.

초기에는 음성 발음 단위인 음소와 오디오 신호에 정렬된 음소 데이터를 오디오 신호와 함께 사용하여 음성 신호를 분리하는 기술이 제안되었다[13]. 이와 같은 기술은 오디오 신호만 사용하는 음성 분리 기술보다 높은 음성 분리 성능을 달성하였다. 하지만, 정렬된 음소 데이터 수집에는 음성 발음에 대한 전문적 지식이 요구되기 때문에 큰 비용이 소요

될 수 있다. 이와 같은 문제를 해결하기 위해 동적 시간 워핑(dynamic time warping; DTW) 기법을 활용하여 음소 시퀀스(sequence) 와 오디오 신호 사이의 정렬을 자동으로 맞추어 학습하는 연구가 있었다. 더 나아가 DTW와 attention 메커니즘을 혼합하여 개선된 성능을 보여주는 연구도 존재한다[14]. 이와 같은 선행 연구들은 오디오 신호와 음소 시퀀스의 정렬 이후 신경망을 적용하여, 분리된 음성 신호를 취득한다. 특히, 트랜스포머 구조를 가진 SepFormer는 지역 및 전역 정보를 반영할 수 있는 dual-path 구조를 음성 분리 신경망에 적용하여 우수한 성능을 달성했다.

하지만, 기존 음성 분리 기술 연구들은 크게 두 가지 한계점이 존재한다. 첫번째로 음소 시퀀스 사용에 있어 각 음소들의 연관성이 낮기 때문에 텍스트 데이터에서 콘텐츠의 문맥 정보를 취득하는데 한계가 존재한다. 음소 시퀀스는 텍스트 인코더라는 신경망을 거치면서 문맥 정보를 반영할 수 있지만, 대규모 텍스트 데이터로 학습된 언어 모델과 비교하여 문맥에 대한 정보를 표현하는데 한계가 존재한다[15]. 두번째로 음소 시퀀스와 오디오 신호가 정렬된 이후 신경망을 거치면서 정렬된 정보가 희석될 수 있다. 이는 신경망의 레이어가 많아질수록 정보가 희석될 수 있으며[16], 음성 분리 성능을 올리기 위해 레이어 수를 증가시키는데 한계로 작용할 수 있다. 이와 같은 문제로 기존 연구들은 텍스트 데이터를 활용한 음성 분리 기술 성능 향상에 한계가 존재할 수 있다.

이에 따라 본 논문에서는 주어진 콘텐츠의 텍스트 데이터의 문맥 정보를 활용하기 위해 대규모 텍스트 데이터로 학습된 텍스트 임베딩 모델과 트랜스포머 기반 음성 분리 신경망 모델을 결합한 텍스트-오디오 기반 음성 분리 기술을 제안한다. 제안된 음성 분리 기술에서 언어 임베딩 모델은 주어진 텍스트 데이터에서 문맥 정보 및 텍스트 토큰의 정보를 추출하기 위해 활용되며, 트랜스포머 기반의 음성 분리 신경망 모델은 지역 및 전역 정보를 반영할 수 있는 모듈에 cross-attention이 적용되어 텍스트 토큰과 오디오 신호의 정렬을 맞출 수 있도록 활용된다. 제안된 텍스트-오디오 기반 음성 분리 신경망 모델은 음성 신호와 배경음악이 혼합된 시뮬레이션 데이터에서 음성 분리 평가 지표를 활용하여 평가된다.

Ⅱ. 관련 연구

제안된 텍스트-오디오 기반 음성 분리 기술은 텍스트 시퀀스에서 문맥 정보 및 텍스트 토큰 추출을 위해 언어 임베딩 모델이 분리 신경망 모델의 인코더로 사용된다. 그리고, cross-attention은 텍스트 시퀀스와 오디오 신호의 정렬을 맞추기 위해 사용되며, 이와 같은 cross-attention은 트랜스포머 기반 음성 분리 모델의 dual-path 구조에 적용되어 오디오 신호에서 음성 신호를 분리한다. 즉, 이 장에서는 텍스트-오디오 기반 음성 분리를 위한 언어 임베딩 모델, cross-attention, 트랜스포머 기반 음성 분리 신경망 모델에 대해 살펴본다.

1. 언어 임베딩 모델

텍스트 데이터의 표현, 즉, 언어 임베딩 추출을 위해 사용되는 신경망 모델 구조는 일반적으로 인코더, 디코더, 인코더-디코더와 같이 3가지로 구분된다[17]. 인코더 기반 모델은 텍스트 데이터 전체 문맥을 양방향으로 처리하여, 텍스트 토큰 사이의 상호 관계를 분석하여 학습할 수 있도록 구성되어 있다. 디코더 기반 모델은 주로 텍스트 데이터 생성 분야에 특화되어 있으며, 주어진 텍스트 토큰들을 사용하여 다음 텍스트 토큰을 예측하는 방식으로 구성된다. 마지막으로 인코더-디코더 구조는 번역이나 요약과 같이 입력과 출력이 서로 다른 텍스트 시퀀스를 가지는 조건부 텍스트 생성형 분야에서 많이 사용된다.

디코더 또는 인코더-디코더 기반의 언어 모델들은 주어진 문장 전체에 대한 문맥 파악보다는 주어진 문장에서 다음 텍스트 토큰 예측을 위한 구조로 인코더 기반의 언어 모델과 비교하여 문맥 파악에 한계점이 존재할 수 있다. 이에 따라

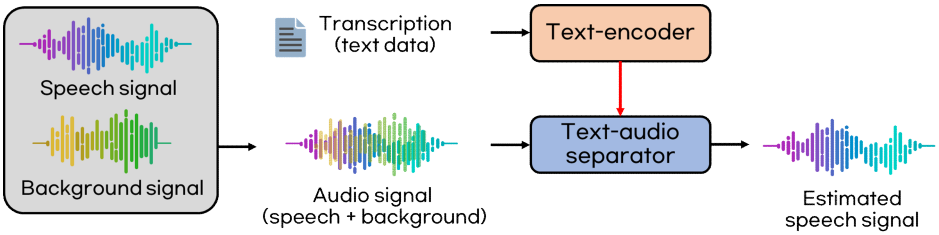

그림 1. 제안된 텍스트-오디오 기반 음성 분리 기술 흐름도

본 논문에서는 콘텐츠의 문맥적 의미 파악에 강점이 있는 인코더 기반의 언어 임베딩 모델을 사용하여, 텍스트-오디오 음성 분리 모델의 인코더로 사용한다.

2. Cross-attention

트랜스포머에는 크게 두 가지 attention 모듈인 self-attention과 cross-attention이 사용된다[18]. 먼저, self-attention은 하나의 입력 시퀀스에서 각 시점의 토큰이 다른 시점의 토큰들과의 관계성을 분석하여, 주어진 입력 시퀀스의 문맥적 연관성을 학습한다. 반면, cross-attention은 서로 다른 두 입력 시퀀스 사이의 관계성을 모델링하기 위해 사용된다. Cross-attention은 하나의 입력 시퀀스에서 쿼리(query)를 생성하고 다른 입력 시퀀스에서 키(key)와 값(value)을 생성하여, 두 시퀀스 사이의 관계도를 계산한다[19].

그러므로, 제안된 텍스트-오디오 기반 음성 분리 기술에서는 cross-attention을 활용하여 텍스트 시퀀스와 오디오 신호 사이의 정렬 수행에 활용된다. 즉, 언어 임베딩 벡터는 query로 사용되고, 오디오 신호는 key와 value로 사용되어, 텍스트 토큰이 오디오 데이터의 어떤 시점과 연결되는지 학습하도록 유도한다. 이와 같은 방식은 강제 정렬(forced alignment) 기반 접근 방법처럼 정렬 정보를 필요하지 않으며, 음성 분리 신경망 학습 시 텍스트-오디오 사이의 정렬이 동시에 학습될 수 있다는 장점이 존재한다.

3. 트랜스포머 기반 음성 분리 신경망

트랜스포머는 multi-head attention과 잔차 연결 등의 구조로 구성되어 있으며, 이는 시퀀스 데이터 내 장기 의존성을 효과적으로 모델링할 수 있는 구조이다. 기존 합성곱 신경망 또는 재귀 신경망 기반 모델이 시계열 정보에 대해 수용영역과 누적 오차를 갖지만, 트랜스포머 구조는 입력 시퀀스 데이터 전체를 동시에 처리함으로써 전역적인 문맥 정보를 효과적으로 활용할 수 있다.

또한, 음성 분리 기술에서도 지역 정보와 전역 정보는 주요한 요소이며, dual-path 구조는 이러한 지역/전역 정보 모두를 반영할 수 있는 구조이다. 이와 같은 dual-path 구조는 음성 분리 신경망에 적용되어 좋은 성능을 달성하고 있다. 이와 같은 dual-path 구조와 트랜스포머 기반 모델 구조를 활용한 SepFormer는 입력 오디오 신호를 일정한 길이의 세그먼트로 분할한다. 그다음, 세그먼트 내의 정보를 처리하는 intra-chunk 구조와 세그먼트 간의 전역 정보를 처리하는 inter-chunk 구조를 순차적으로 통과하여 지역 및 전역 시간 정보를 효과적으로 학습할 수 있다.

본 논문에서는 이와 같은 SepFormer 구조에 텍스트 정보를 통합하기 위해 cross-attention 모듈을 활용한다. 구체적으로, intra-/inter-chunk 구조를 SepFormer 구조 중 일부 레이어에 텍스트 임베딩 벡터와 오디오 특징 벡터 간의 cross-attention 연산을 적용한다. 즉, 음성 분리 모델은 단순히 시간-주파수 패턴만을 기반으로 음성을 분리하는 것이 아니라, 주어진 텍스트 데이터와 문맥적으로 일치하는 음성 신호를 선택적으로 강조함을 목표한다.

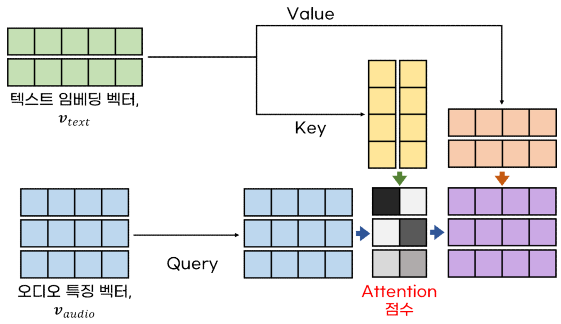

그림 2. 텍스트-오디오 기반 cross-attention 구성도

Ⅲ. 텍스트-오디오 기반 음성 분리 신경망 모델

본 장에서는 제안된 텍스트-오디오 기반 음성 분리 신경망에 대해 기술을 진행한다. 그림 1은 제안된 텍스트-오디오 기반 음성 분리 기술의 흐름도를 나타내며, 음성 신호와 배경음악이 섞여 있는 오디오 신호와 텍스트 데이터가 동시에 음성 분리 모델의 입력 데이터로 사용된다. 다음으로 음성 분리 모델은 입력된 텍스트 및 오디오 데이터를 분석하여 오디오 신호에 포함된 음성 신호를 추출하는 구조이다.

1. 오디오 특징 벡터 추출

제안된 텍스트-오디오 음성 분리 모델에 사용되는 오디오 데이터는 16bits 양자화, 16kHz 샘플링 레이트(sampling rate)를 갖는 오디오 신호 x를 사용한다. 이와 같은 오디오 신호는 10ms hop size와 32ms window size로 short-time Fourier transform(STFT)을 거쳐 스펙트로그램(spectrogram) X로 변환된다. 즉, 시간 도메인의 오디오 신호 x는 시간-주파수 도메인의 오디오 스펙트로그램 X로 변환된다.

추출된 스펙트로그램은 시간-주파수 도메인에서 복소수로 표현되며, 절댓값 연산을 통해 스펙트로그램의 크기로 변환한다. 이러한 스펙트로그램 크기 $|X|$는 텍스트-오디오 기반 음성 분리 모델의 오디오 특징 벡터 v_{audio}로 사용되며, 오디오 신호에서 음성 특징이 존재하는 부분에 대한 마스크 예측에 활용된다.

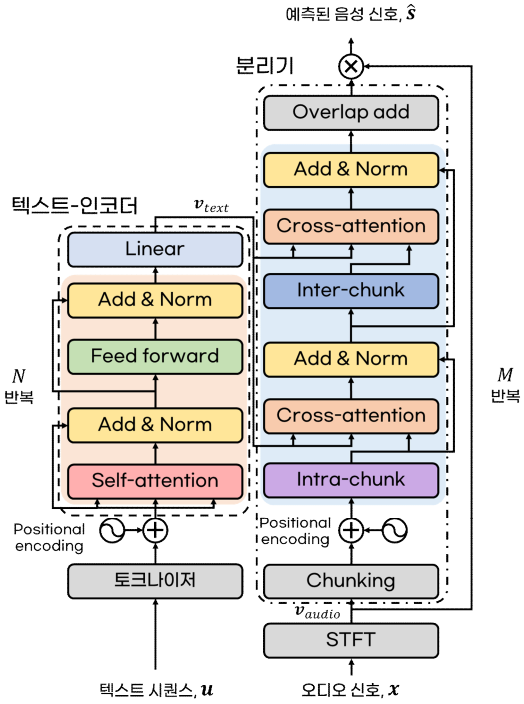

그림 3. 제안된 텍스트-오디오 기반 음성 분리 신경망 모델 구성도

2. 텍스트-오디오 cross-attention 구조

본 논문에서는 텍스트 시퀀스와 오디오 신호 사이의 정렬과 문맥 정보 반영을 위해 cross-attention 구조가 활용된다. 그림 2는 텍스트 임베딩 벡터 v_{text}와 오디오 특징 벡터 v_{audio} 사이의 정렬과 문맥 정보 반영을 위한 구조를 나타낸다. 그림에서 볼 수 있듯이 텍스트 임베딩 벡터는 value와 key로 사용되며, 오디오 특징 벡터는 query로 활용된다.

먼저, key와 query 벡터들의 내적 연산을 통해 attention 점수가 산출되며, 이는 텍스트 토큰이 오디오 신호의 어느 부분과 연관성이 있는지 확률로 표현한다. 다음으로 attention 점수를 이용해 value를 가중합하여, 오디오 특징 벡터에 문맥적으로 일치하는 위치를 강조하여, 텍스트 정보가 보강된 새로운 특징 벡터를 추출한다. 즉, 이와 같은 구조를 통해 텍스트 데이터가 오디오 데이터와의 정렬과 동시에 문맥적 의미를 반영할 수 있는 특징 벡터로 변환되게 된다.

3. 텍스트-오디오 음성 분리 신경망 구조

본 논문에서 제안된 텍스트-오디오 기반 음성 분리 신경망의 자세한 구조는 그림 3과 같다. 그림에서 볼 수 있듯이 제안된 음성 분리 신경망 구조는 텍스트-인코더와 분리기(separator)로 구성되어 있다. 먼저, 텍스트-인코더는 텍스트 임베딩 벡터를 추출하는 언어 임베딩 모델, 오디오-인코더는 오디오 특징 벡터 추출 모듈을 나타낸다. 즉, 텍스트-인코더와 오디오-인코더에서는 v_{audio}와 v_{text} 벡터가 각각 출력되며 cross-attention 모듈의 입력으로 사용된다.

먼저, 텍스트-인코더는 인코더 구조는 일반적인 텍스트 임베딩 모델의 구조로 구성되어 있으며, 텍스트 시퀀스 u는 토크나이저(tokenizer)를 거쳐 토큰 단위 변환되어 텍스트 임베딩 모델에 입력된다. 텍스트-인코더 내부의 self-attention으로 구성된 블록은 N번 만큼 반복되어 구성되며, 이는 텍스트 임베딩 모델마다 다른 값을 갖는다. 이와 같이 텍스트 시퀀스로부터 텍스트-인코더를 거쳐 생성된 v_{text} 벡터는 cross-attention 모듈의 key와 value로 사용된다.

다음으로 분리기는 SepFormer의 구조를 기반하고 있으며, intra-/inter-chunk로 구성된 모듈이 M번 만큼 박복되는 구조로 구성되어 있다. 분리기에서 각각의 intra-/inter-chunk 구조 다음에는 기존 self-attention 구조 대신 본 논문에서 제안한 cross-attention 구조가 적용되어 있다. 즉, 첫 번째 intra-chunk 구조 다음으로 구성된 cross-attention 구조에서는 텍스트-오디오 데이터의 지역적 표현을 학습한다. 이와 유사하게 inter-chunk 구조 다음으로 구성된 cross-attention 구조에서는 텍스트-오디오 데이터의 전역적 표현을 학습한다. 즉, 분리기 내부 cross-attention 모듈에서 지역/전역적인 문맥 정보를 기반한 텍스트-오디오 정렬을 학습시키는 구조를 가지게 된다.

구성된 모델은 음성 분리 신경망 학습에 주로 사용되는 scale-invariant signal-to-distortion ratio(SI-SDR) 손실함수 $\mathcal{L}_{si-sdr}$[20]와 스펙트로그램 손실함수 $\mathcal{L}_{spec}$[21]를 통해 학습되며, 손실함수는 각각 수식 (1)과 (2)와 같이 정의된다.

$$\mathcal{L}_{spec} = \| \boldsymbol{S} - \tilde{\boldsymbol{S}} \|_2^2 \quad (2)$$

수식에서 $\boldsymbol{s}$와 $\tilde{\boldsymbol{s}}$는 각각 음성 신호와 예측된 음성 신호를 나타내고, $\boldsymbol{S}$와 $\tilde{\boldsymbol{S}}$는 각각 음성 신호와 예측 음성 신호의 스펙트로그램 표현을 나타낸다. 즉, $\mathcal{L}_{si-sdr}$는 시간 도메인에서 손실함수를 나타내고, $\mathcal{L}_{spec}$는 시간-주파수 도메인에서 손실함수를 나타낸다.

제안된 텍스트-오디오 기반 음성 분리 모델 과정에서 텍스트-인코더의 모델 가중치는 고정한 채 분리기의 가중치만 업데이트한다. 그림 3에서 표현된 눈 모양은 모델 가중치 고정을 의미하고 불꽃 모양은 모델 가중치 학습을 의미한다.

Ⅳ. 실험 및 성능평가

본 장에서는 제안된 텍스트-오디오 기반 음성 분리 신경망에 시뮬레이션 데이터를 이용한 실험 환경 및 성능평가에 관해 기술한다.

1. 데이터 구성

음성 분리 성능평가를 위해서 본 논문에서는 음성 인식 분야에 주로 사용되는 LibriSpeech 데이터[22]를 음성 신호 데이터로 활용하고, 오디오 음원 분리 분야 등에 주로 사용되는 MUSDB18 데이터[23]를 배경음악 데이터로 활용한다. LibriSpeech 데이터셋은 16bits 양자화, 16kHz로 샘플링되어 있으며, 음성 신호와 대응하는 텍스트 데이터가 약 960시간 포함되어 있다. MUSDB18 데이터셋은 32bits 양자화, 22.05kHz로 샘플링 되어 있으며, 해당 데이터에는 보컬, 드럼, 베이스, 기타로 총 4가지 종류의 음원을 포함하고 있다.

두 데이터셋은 서로 다른 샘플링 레이트를 가지고 있으므로, MUSDB18 데이터셋을 16kHz로

다운 샘플링하고 16bits로 양자화하여, LibriSpeech 데이터셋과 동일한 샘플링률로 다운 샘플링한다. 그리고, 음성 신호와 배경음악 신호의 signal-to-noise ratio(SNR)은 0dB부터 10dB 사이에서 균등하게 선택하여 음성 신호와 배경음악 신호를 믹싱(mixing)한다. 본 실험에서 사용되는 학습데이터는 LibriSpeech 데이터셋의 'train-clean-100', 'train-clean-360', train-other-500' 데이터에 MUSDB18 데이터셋의 'train'에 포함된 데이터와 믹싱한다. 그리고, 검증데이터는 LibriSpeech의 'dev-clean'과 'dev-other'에 MUSDB18의 'train'에 포함된 음원과 믹싱한다. 마지막으로 평가 데이터는 LibriSpeech의 'test-clean'과 'test-other'에 MUSDB18의 'test'에 포함된 음원과 믹싱한다.

추가로 실험에 사용하는 데이터셋을 크게 2가지로 분류한다. 데이터셋 하나는 배경음악 데이터에서 MUSDB18에 존재하는 보컬 데이터를 제외하고, 나머지 데이터셋은 보컬 데이터를 모두 포함하여 학습 및 평가를 진행한다.

2. 실험 환경 설정

텍스트 임베딩 모델은 대규모 영어 텍스트로 사전 학습된 BERT[24], RoBERTa[25], ELECTRA[26] 모델들을 활용하여, 텍스트-오디오 음성 분리 신경망 모델을 학습한다. 구체적으로 Hugging Face에 공개된 [1] bert-base-uncased, [2]roberta-base, [3] electra-base-discriminator 3가지 모델을 각각 사용하여, 각 모델에 따른 성능평가를 진행한다.

제안된 텍스트-오디오 기반 음성 신경망 모델은 코사인 스케줄 기반 애열-재시작 기법[27]에 따라 학습률이 변화하며 학습이 진행된다. 그리고, Sepformer에 포함된 퍼뮤트 연산의 특성을 반영하기 위해 미니 배치 크기는 1로 설정하여 학습을 진행한다. 마지막으로, Sepformer의 분리

[1]https://huggingface.co/google-bert/bert-base-uncased
[2]https://huggingface.co/FacebookAI/roberta-base
[3]https://huggingface.co/google/electra-base-discriminator

$$\mathcal{L}_{si-sdr} = \frac{\left\| \frac{\langle \tilde{s}, s \rangle}{\| s \|^2} s \right\|^2}{\left\| \tilde{s} - \frac{\langle \tilde{s}, s \rangle}{\| s \|^2} s \right\|^2} \quad (1)$$

기의 반복횟수, M은 8로 설정한다.

본 실험에서 음성 분리의 성능평가 지표로는 signal-to-distortion(SDR), signal-to-artifact (SAR), signal-to-interference(SIR)을 이용하여 음성 분리도에 대한 성능[28]을 측정한다. 그리고, perceptual evaluation of speech quality(PESQ)[29]와 short-time objective intelligibility(STOI)[30]를 사용하여 음성 품질에 대한 성능을 측정한다.

3. 실험 결과

기존 음성 분리 성능 비교를 위해 텍스트 정보 없이 학습된 SepFormer 기반의 음성 분리 모델 (SepFormer(baseline))과 비교를 진행한다. 또한, SepFormer에 텍스트와 오디오 사이의 정렬을 맞추는 DTW와 attention 메커니즘을 결합한 기술[14]을 적용한 음성 분리 모델 (텍스트-오디오 SepFormer(DTW-att.))과 재귀 신경만 기반 음

표 1. 보컬 음원이 포함되지 않은 데이터셋에서의 제안된 음성 분리 기술의 성능평가

	SDR(dB)	SAR(dB)	SIR(dB)	PESQ	STOI
SepFormer (baseline)	16.57	19.28	16.91	3.24	0.91
텍스트-오디오 Phoneme-RNN	15.89	18.72	15.72	3.17	0.88
텍스트-오디오 SepFormer (DTW-att.)	18.68	21.90	19.15	3.31	0.93
제안된 텍스트-오디오 Sepformer (BERT)	19.41	22.52	20.45	3.50	0.95
제안된 텍스트-오디오 Sepformer (RoBERTa)	19.37	22.34	20.42	3.48	0.95
제안된 텍스트-오디오 Sepformer (ELECTRA)	19.25	22.21	20.28	3.47	0.95

성 분리 모델에 음소 정렬을 맞추기 위한 DTW를 음성 분리 모델(텍스트-오디오 Phoneme-RNN)[31]과 비교를 진행한다. 그리고, 제안된 텍스트-오디오 기반 음성 분리 모델에 대한 평가와 함께 텍스트 임베딩 모델에 대한 의존성이 없음을 확인하기 위해 서로 다른 텍스트 임베딩 모델들(BERT, RoBERTa, ELECTRA)을 텍스트-인코더에 적용하여 실험 결과를 비교한다.

표 1은 보컬 음원을 포함하지 않은 배경음악 환경에서의 음성 분리 성능을 나타낸다. 표에서 볼 수 있듯이 본 실험에서 기준이 되는 SepFormer는 음성 분리 지표에서 좋은 성능을 나타내고 있으며, 텍스트 정보를 사용하는 Phoneme-RNN 음성 분리 모델보다 좋은 성능을 나타내고 있다. 즉, 텍스트 정보를 활용하는 재귀 신경망 기반 음성 분리 모델이 오디오 데이터만 사용하는 트랜스포머 기반 음성 분리 모델보다 낮은 성능을 보여주며, 이는 트랜스포머의 강력한 음성 분리 성능을 보여준다. 그리고, DTW와 attention 메커니즘을 적용한 음성 분리 모델은 텍스트 정보를 활용하기 때문에, 오디오 데이터만으로 학습된 SepFormer보다 좋은 성능을 나타내고 있다. 더 나아가 제안된 텍스트-오디오 기반 음성 분리 신경망 모델은 기본 SepFormer보다 향상된 음성 분리 지표를 나타내고 있으며, DTW와 attention 메커니즘을 적용한 텍스트-오디오 기반 음성 분리 모델보다 좋은 성능을 나타내고 있다.

특히, 음성 분리도를 나타내는 SDR, SAR, SIR은 큰 폭으로 향상됨을 확인할 수 있다. 이러한 결과는 제안된 텍스트-오디오 기반 음성 분리 모델이 문맥적인 정보를 잘 반영하고 텍스트-오디오 사이의 정렬에서 상대적으로 우수한 성능을 달성한 것으로 판단된다. 또한, 제안된 텍스트-오디오 기반 음성 분리 기술은 텍스트 임베딩 모델의 종류에 상관없이 일관되게 성능이 개선됨을 확인할 수 있다.

표 2는 보컬 음원이 포함된 배경음악 환경에서

표 2. 보컬 음원이 포함된 데이터셋에서의 제안된 음성 분리 기술의 성능평가

	SDR(dB)	SAR(dB)	SIR(dB)	PESQ	STOI
SepFormer (baseline)	8.62	10.05	9.02	2.87	0.82
텍스트-오디오 Phoneme-RNN	11.85	11.61	12.06	2.91	0.83
텍스트-오디오 SepFormer (DTW-att.)	13.25	12.52	13.42	2.96	0.84
제안된 텍스트-오디오 Sepformer (BERT)	16.54	14.05	16.61	3.14	0.86
제안된 텍스트-오디오 Sepformer (RoBERTa)	16.42	13.91	16.42	3.04	0.85
제안된 텍스트-오디오 Sepformer (ELECTRA)	16.28	13.97	16.37	3.07	0.85

각 모델들의 음성 분리 성능을 나타낸다. 표에서 볼 수 있듯이 전반적인 성능 하락이 발생하고 있으며, 이와 같은 원인은 배경음악에 포함된 보컬 음원은 음성 신호와 유사한 특성을 나타내기 때문이다. 특히, 텍스트 정보 없이 학습된 SepFormer는 특성이 유사한 보컬 음원과 음성 신호에 대한 분리가 잘 진행되지 않아 상당히 낮은 음성 분리 성능을 나타내었다. 이와 같은 환경으로 인해 RNN-Phoneme 기반 음성 분리 모델은 표 1과 다르게 SepFormer(baseline)보다 높은 성능을 나타내고 있다.

본 논문에서 제안한 텍스트-오디오 기반 음성 분리 기술은 오디오 신호와 텍스트 정보를 활용하여 음성 분리를 진행하였기 때문에 음성 분리 평가 지표들이 상대적으로 큰 폭으로 개선됨을 확인할 수 있다. 표 1과 유사하게 SepFormer (DTW-att.)와 Phoneme-RNN 기반 음성 분리 모델 보다 좋은 성능을 달성하고 있다. 또한, 다양한 텍스트 임베딩 모델을 텍스트-인코더로 적용해도 일관된 성능을 나타냄을 확인할 수 있다.

Ⅴ. 결론

본 논문에서는 텍스트 데이터를 활용한 텍스트-오디오 기반 음성 분리 기술을 제안하였다. 제안된 모델은 대규모 텍스트 데이터로 사전 학습된 언어 임베딩 모델을 사용하여 문맥적 의미와 토큰 단위 정보를 추출하고, SepFormer기반 음성 분리 신경망 모델에 cross-attention을 적용하여 텍스트와 오디오 사이의 정렬을 학습하였다. 시뮬레이션 데이터를 사용한 성능평가에서 제안된 음성 분리 모델은 오디오 데이터만 활용한 SepFormer 및 텍스트-오디오 데이터를 활용한 DTW-attention 기반 음성 분리 기술보다 SDR, SAR, SIR, PESQ, STOI 지표에서 모두 우수한 성능을 나타냈다. 특히, 보컬 음원이 포함된 배경음악 환경에서도 텍스트 데이터를 활용함으로써 기존 방법 대비 우수한 성능을 달성했다. 이와 같은 결과는 학술적으로 텍스트-오디오 기반 음성 분리 분야에서 텍스트와 오디오 사이의 정렬에서 이진 정렬보다 확률적 정렬의 우수함을 확인하였다. 이와 같은 텍스트-오디오 기반 음성 분리 기술을 실제 환경에서 콘텐츠 리믹싱 및 고품질 오디오 편집에 활용되어 사용자에게 큰 만족감을 제공할 것으로 기대한다.

제안된 텍스트-오디오 기반 음성 분리 모델은 종래 기술보다 우수한 음성 분리 성능을 달성하였다. 하지만, PESQ가 3.0을 근소하게 넘는 수치로 양호한 음질이나 일부 청취자가 품질 저하를 인지할 수 있다. 즉, 청취자에게 고품질 음성을 제공하기 위해서는 성능 개선이 요구된다. DTW 기반 오디오-텍스트 정렬은 이산 정렬, attention 메커니즘은 확률적 정렬에 속한다고 해석될 수 있으며, 본 논문의 실험 결과는 확률적 정렬이 텍스트-오디오 기반 음성 분리에 적합함을 암시한다. 확률적 정렬에서 우수한 성능을 나타내는 monotonic 정렬을 사용한 연구를 수행하면, 향후 우수한 성능을 나타내는 텍스트-오디오 기반 음성 분리 성능을 달성할 것으로 기대한다.

REFERENCES

[1] S. Alharbi, et al., “Automatic speech recognition: Systematic literature review,” *IEEE Access*, vol. 9, pp. 131858 - 131876, 2021.

[2] Y. Kumar, A. Koul, and C. Singh, “A deep learning approaches in text-to-speech system: A systematic review and recent research perspective,” *Multimedia Tools and Applications*, vol. 82, no. 10, pp. 15171 - 15197, 2023.

[3] D. Michelsanti, et al., “An overview of deep-learning-based audio-visual speech enhancement and separation,” *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 29, pp. 1368 - 1396, 2021.

[4] D. Wang and J. Chen, “Supervised speech separation based on deep learning: An overview,” *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 26, no. 10, pp. 1702 - 1726, 2018.

[5] J. Pons, et al., “Remixing music using source separation algorithms to improve the musical experience of cochlear implant users,” *J. Acoust. Soc. Am.*, vol. 140, no. 6, pp. 4338 - 4349, 2016.

[6] K. Tan, et al., “Audio-visual speech separation and dereverberation with a two-stage multimodal network,” *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 3, pp. 542 - 553, 2020.

[7] P. Huang, et al., “Deep learning for monaural speech separation,” in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[8] W. H. Nam, et al., “AI 3D immersive audio codec based on content-adaptive dynamic down-mixing and up-mixing framework,” in *Proc. Audio Eng. Soc. Convention 151*, 2021.

[9] S. Choi, et al., “Blind source separation and independent component analysis: A review,” *Neural Information Processing - Letters and Reviews,* vol. 6, no. 1, pp. 1 - 57, 2005.

[10] Y. Luo and N. Mesgarani, “Conv-TasNet: Surpassing ideal time - frequency magnitude masking for speech separation,” *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 27, no. 8, pp. 1256 - 1266, 2019.

[11] Y. Luo, Z. Chen, and T. Yoshioka, “Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation,” in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

[12] C. Subakan, et al., “Attention is all you need in

speech separation," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2021.
[13] G. Meseguer-Brocal and G. Peeters, "Content based singing voice source separation via strong conditioning using aligned phonemes," arXiv preprint arXiv:2008.02070, 2020.
[14] K. Schulze-Forster, et al., "Phoneme level lyrics alignment and text-informed singing voice separation," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 29, pp. 2382 - 2395, 2021.
[15] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," arXiv preprint arXiv:1908.10084, 2019.
[16] R. Xiong, et al., "On layer normalization in the transformer architecture," in *Proc. Int. Conf. Machine Learning (ICML)*, PMLR, 2020.
[17] T. Lin, et al., "A survey of transformers," *AI Open*, vol. 3, pp. 111 - 132, 2022.
[18] A. Vaswani, et al., "Attention is all you need," in Advances in *Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
[19] D. Bahdanau, et al., "End-to-end attention-based large vocabulary speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2016.
[20] X. Hao, et al., "FullSubNet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2021.
[21] U. Isik, et al., "PocoNet: Better speech enhancement with frequency-positional embeddings, semi-supervised conversational data, and biased loss," arXiv preprint arXiv:2008.04470, 2020.
[22] V. Panayotov, et al., "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2015.
[23] Z. Rafii, et al., "The MUSDB18 corpus for music separation," 2017. [Online]. Available: https://sigsep.github.io/datasets/musdb.html
[24] J. Devlin, et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, pp. 4171 - 4186, 2019.
[25] Y. Liu, et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.
[26] L. Clark, et al., "ELECTRA: Pre-training text encoders as discriminators rather than generators," arXiv preprint arXiv:2003.10555, 2020.
[27] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," arXiv preprint arXiv:1608.03983, 2016.
[28] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 14, no. 4, pp. 1462 - 1470, 2006.
[29] A. W. Rix, et al., "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 749 - 752, 2001.
[30] C. H. Taal, et al., "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 4214 - 4217, 2010.
[31] K. Schulze-Forster, C. S. Doire, G. Richard, and R. Badeau, "Joint phoneme alignment and text-informed speech separation on highly corrupted speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)* pp. 7274 - 7278, 2020.

저 자 소 개

이건우(정회원)
2017년 전남대학교 전자컴퓨터공학부 학사 졸업.
2019년 광주과학기술원 전기전자컴퓨터공학부 석사 졸업.
2004년 광주과학기술원 AI대학원 박사 졸업.
<주관심분야 : 음성 잡음 제거, 음성 인식, 자연어 처리>