Spleeter, BSRNN, Demucs, Open-Unmix 딥러닝 기반 음원 소스 분리 기술의 종합적 고찰

오주원, 이재호*
덕성여자대학교, *덕성여자대학교

2000violet@duksung.ac.kr, *izeho@duksung.ac.kr

A Comprehensive Survey of Music Source Separation Model Based on Deep Learning: Spleeter, BSRNN, Demucs, Open-Unmix, D3Net

JooWon OH, Jaeho Lee*
Duksung Women's Univ., *Duksung Women's Univ.

요 약

음원 소스 분리는 하나의 혼합된 사운드—대개 스테레오 형태로 공간적으로 혼합된 신호—로부터 보컬, 드럼, 베이스, 피아노 및 기타 반주 등 개별 악기 소스를 복원하는 음악 신호 처리의 핵심 작업이다. 최근 딥러닝 기술의 급속한 발전에 따라 다양한 신경망 기반 소스 분리 모델이 제안되었으며, 실제 음악 응용에도 폭넓게 활용되고 있다. 본 논문에서는 대표적인 딥러닝 기반 음악 소스 분리 모델인 Spleeter, Open-Unmix, Demucs, BSRNN, D3Net 의 구조와 성능을 종합적으로 비교·분석하였다. 각 모델은 U-Net, BiLSTM, 파형 기반 네트워크, 밴드 분할 RNN, 다중 해상도 컨볼루션 등 서로 다른 아키텍처적 특성과 처리 방식을 갖고 있으며, 시간-주파수 표현 처리 방식, 수용 영역 확장 전략, stem 분리 구성 등에 따라 차별점을 가진다. 특히 Spleeter 는 유일하게 5 개의 stem(보컬, 드럼, 베이스, 피아노, 기타) 분리를 지원하며, 나머지 모델들은 4 stem 분리를 기본으로 한다. 각 모델의 구조, 입력 및 출력 형태(STFT 또는 waveform), 훈련 전략 등을 종합적으로 분석함으로써, 효율적이고 강건한 음악 소스 분리 시스템 설계에 필요한 핵심 인사이트를 제공한다.

Ⅰ. 서 론

음원 소스 분리는 멀티 채널일 수 있는 관찰된 믹싱 소스, 즉 혼합된 오디오로부터 개별 구성 요소 사운드를 복구하는 신호 처리 작업이다. 음악 신호 처리에서 복원 대상인 소스는 보통 보컬, 드럼, 베이스, 기타 등 다양한 악기 소리와 일치하며, 이러한 믹싱은 대부분 입체 음향으로 구성된다[1]. 최근 음악 생성, 리마스터링, 카라오케, 보컬 제거, 오디오 분석 등 다양한 분야에서 음원 소스 분리 기술이 활발히 응용되고 있으며, 특히 딥러닝 기반의 접근 방식이 주류를 이루고 있다.

기존의 전통적 방법론은 주로 통계 기반 필터링이나 NMF(Non-negative Matrix Factorization)와 같은 기술을 사용했지만, 최근에는 시간-주파수 특성 및 위상 정보를 정밀하게 학습할 수 있는 딥러닝 기반 모델들이 뛰어난 성능을 보이고 있다. 이에 따라 본 논문에서는 대표적인 딥러닝 기반 음원 소스 분리 모델인 Spleeter, Demucs, Open-Unmix, BSRNN, D3Net 을 중심으로 구조적 차이와 분리 성능을 비교 분석하고, 각 모델의 적용 가능성과 특성을 고찰한다.

Ⅱ. 본론

Spleeter[2]는 Deezer 에서 개발한 음악 소스 분리 도구로, 빠른 처리 속도와 우수한 성능을 특징으로 한다. 사용자는 음악 파일을 보컬과 반주 (2stem), 혹은 보컬, 드럼, 베이스, 기타(4stem), 그리고 피아노까지 포함한 5stem 까지 분리 가능하다. Spleeter 는 Tensorflow 기반으로 구현되었으며, 사전 학습된 모델을 제공한다. Adam Optimizer 를 통해 훈련을 최적화했으며, 단일 GPU 에서 약 일주일 정도 훈련한 모델이다.

모델로는 12-layer U-Net 기반의 CNN 구조를 채택하고 있다. Encoder 와 Decoder 각각 6 개의 Layer 로 구성되어 있으며, 이 두 개 사이에 Skip Connection 이 사용되어 encoder 의 feature map 을 decoder 에 직접 연결하여 정보 손실을 최소화한다. 입력 믹스의 Spectrogram 에서 각 소스의 soft-mask 를 예측한 후, 이를 통해 원래의 각 소스를 재구성한다. 이 mask 를 혼합된 Spectrogram 에 곱하여 분리된 각 소스의 Spectrogram 을 생성한다.

Open-Unmix[3]는 프랑스 INRIA 와 Sony Europe 이 협력하여 개발한 음원 소스 분리 모델로 SiSEC 평가 캠페인에서 상위권 성능을 기록하였다. Spectrogram 기반의 모델로, STFT(Short-Time Fourier Transform)으로 오디오를 변환한 후, 각 소스에 대한

마스크를 예측하고, 예측된 마스크를 적용하여 ISTFT 로 원래의 파형을 복원하는 방식이다.

3 개의 계층으로 이루어진 Bi-directional LSTM 를 사용하여 시퀀스 정보를 처리한 후 Linear Projection Layer 를 통해 소스 별 mask 를 출력한다. 그리고, 예측된 mask 를 기존 Spectrogram 에 곱한 후 ISTFT 를 수행한다.

	입력	출력	모델 구조
Spleeter	STFT Spectrogram	각 소스별 마스크	U-Net (12-layer CNN)
Open-Unmix	STFT Spectrogram	각 소스별 마스크	3-layer BiLSTM + Linear
BSTNN	STFT Spectrogram	각 소스별 마스크	Band-Split + BiLSTM
D3Net	STFT Spectrogram	각 소스별 마스크	DenseNet + Multi-dilated Convolution
Demucs	Raw waveform	waveform	U-Net + 2-layer BiLSTM

BSRNN[4]은 주파수 영역에서 작동한 음악 소스 부리 모델로, 기존 모델들이 다른 audio task 에서 가져온

구조를 사용하는 데 반해 음악 신호의 고유 특성에 최적화된 구조를 설계한 것이 특징이다. 이 모델은 입력 오디오의 STFT Spectrogram 을 여러 subbands 로 나눈 후, 각각의 band 에 대해 순차적으로 RNN 처리를 수행한다. 최종적으로 이를 통해 time-frequency mask 를 예측하여 소스를 분리한다. 특히 band 의 개수와 대역폭을 악기별 특성에 따라 조정 가능하다는 점이 주요 강점이다.

이 모델은 사전 정의된 Band Split Module 의 주파수 대역폭에 따라 STFT 를 여러 band 로 분할한다. 이후 Band and Sequence Modeling 의 BLSTM 을 사용해 각 band 를 sequence 와 band 차원에서 처리한다. Mask Estimation Module 은 각 band 에 대해 개별적으로 MLP 를 통해 complex-valued T-F mask 를 예측한다. 8 개의 GPU 를 병렬로 배치하여 Adam Optimizer 를 통해 100epoch 를 학습하였고, 추가로 semi-supervised fine-tuning 을 도입하여 성능을 높였다.

	Stem	Vocal	Drum	Bass	Piano	Other
(A)	5	O	O	O	O	O
(B)	4	O	O	O	X	O
(C)	4	O	O	O	X	O
(D)	4	O	O	O	X	O
(E)	4	O	O	O	X	O

D3Net[5]은 Sony 에서 제안한 음원 소스 분리 모델로, 기존 CNN 기반 모델들이 겪은 Receptive field 의 성장 한계와 해상도 손실 문제를 동시에 해결하기 위해 고안되었다. 이 모델은 DenseNet 구조를 기반으로 하면서, 한 층에서 여러 Dilation Factor 를 적용할 수 있는 Multidilated Convolution Layer 를 도입한 것이 특징이다. DenseNet 구조 위에 Dilation 연산을 효과적으로 적용하기 위한 다층 블록 구조로, 지역 및 전역 정보를 동시에 추출하여 다중 해상도 처리를 가능하게 한다.

Demucs(Deep Extractor for Music Sources)는 Facebook AI Research 에서 제안한 음악 소스 분리 모델이다. 이는 기존 음원 소스 분리 모델들의 취약점을 극복하고자 설계되어, 입력 오디오를 STFT Spectrogram 변환 후 분리를 수행하는 것과 달리 Demucs 는 처음부터 끝까지 waveform 단위로 처리한다. 이는 위상 정보 손실과 역변환에 따른 artifact 를 줄이는 것이 가능하다.

이는 U-Net 기반 구조에 GLU(Gated Linear Unit) 활성화 함수와 Bi-directional LSTM 을 결합한 구조로, 총 6 개의 Encoder-Decoder 계층을 보유하며, 각 계층은 Skip Connection 으로 연결되어 고해상도 정보를 보존한다. Demucs 의 강점 중 하나로는 Rescaling 학습 전략을 도입하여 학습 안정성과 성능을 향상시킨다는 것이다. 16 개의 V100 GPU 에서 학습하였으며, 시간 이동, 채널 스왑 등과 같은 방식의 데이터 증강을 수행하여 학습에 사용한다.

Spleeter 는 사전 학습된 모델 구성을 통해 2, 4, 5 stems 구성이 가능하며, 보컬(Vocals), 드럼(Drums), 베이스(Bass), 피아노(Piano), 기타 악기(Other Accompaniment 의 다섯 가지 소스 분리를 지원한다. 모델은 각 소스에 대해 T-F 마스크를 예측한 후, Spectrogram 기반 복원을 통해 개별 신호를 분리한다.

Open-Unmix 는 기본적으로 4 stems 구조를 기반으로, 보컬(Vocals), 드럼(Drums), 베이스(Bass), 기타(Other) 소스를 분리한다. 각 악기에 대해 개별적으로 훈련된 Bi-directional LSTM 네트워크를 사용하며, 복원된 Spectrogram 을 기반으로 각 악기 신호를 추정한다.

BSRNN 은 기본적으로 4 stems (보컬, 드럼, 베이스, 기타)의 분리를 수행하며, 악기별로 주파수 대역을 나누어 처리하는 Band-split 전략을 적용한다. 각 소스에 대해 sub-band 및 시간 정보를 동시 학습하는 RNN 기반 구조를 통해 높은 분리 성능을 제공한다.

Demucs 는 파형 기반 분리 모델로, 사전 훈련된 설정에서는 4 stems (보컬, 드럼, 베이스, 기타 또는 기타 악기)를 분리한다. 확장된 구성에서는 사용자 정의에 따라 더 많은 소스를 처리할 수 있으며, 모델 구조 상 멀티 채널 오디오 입력을 그대로 처리하여 위상 정보를 보존한다.

D3Net 은 STFT 기반의 소스 분리 모델로, 4 stems (보컬, 드럼, 베이스, 기타)의 분리를 목표로 설계되었으며, 각 소스마다 별도의 네트워크가 학습된다. 모델은 Multidilated Convolution 과 DenseNet 을 결합한 구조를 사용하여 다양한 해상도에서 특징을 효과적으로 추출한다.

	Train Dataset
Spleeter	Deezer's Internal Dataset
Open-Unmix	MUSDB18[6], 10000hours Speech
BSTNN	MUSDB8-HQ[7],1750 Songs
D3Net	MUSDB18[6]
Demucs	3500 Songs

Ⅲ. 결론

. 본 논문에서는 대표적인 딥러닝 기반 음원 소스 분리 모델인 Spleeter, Open-Unmix, Demucs, BSRNN, D3Net 의 구조적 특징과 분리 성능을 비교·분석하였다. 각 모델은 처리 방식(STFT 기반 vs. waveform 기반), 네트워크 아키텍처, stem 구성 지원 등에 따라 고유한 강점을 보인다. 특히 BSRNN 과 D3Net 은 멀티 해상도 정보 처리 측면에서 우수한 성능을 보이며, Spleeter 는 실시간 처리와 다중 stem 지원에서 강점을 가진다. 본 비교는 다양한 응용 환경에서 적합한 모델 선택을 위한 실질적인 기준을 제공한다. 향후에는 사용자 목적에 맞는 하이브리드 접근 방식도 주요 연구 방향이 될 수 있다.

참 고 문 헌

[1] Liutkus, A., Durrieu, J. L., Daudet, L., & Richard, G. (2013, July). An overview of informed audio source separation. In 2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS) (pp. 1-4). IEEE.

[2] Hennequin, R., Khlif, A., Voituret, F., & Moussallam, M. (2020). Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, *5*(50), 2154.

[3] Stöter, F. R., Uhlich, S., Liutkus, A., & Mitsufuji, Y. (2019). Open-unmix-a reference implementation for music source separation. *Journal of Open Source Software*, *4*(41), 1667.

[4] Yu, J., Luo, Y., Chen, H., Gu, R., & Weng, C. (2022). High fidelity speech enhancement with band-split rnn. *arXiv preprint arXiv:2212.00406*.

[5] Takahashi, N., & Mitsufuji, Y. (2020). D3net: Densely connected multidilated densenet for music source separation. *arXiv preprint arXiv:2010.01733*.

[6] https://paperswithcode.com/dataset/musdb18

[7] https://paperswithcode.com/dataset/musdb18-hq