논문 20-45-05-14 The Journal of Korean Institute of Communications and Information Sciences '20-05 Vol.45 No.05
https://doi.org/10.7840/kics.2020.45.5.855

컨볼루션 신경망 기반 모노 채널 음악-대사 음원 분리 기술을 이용한 방송물 배경 음악 식별

김 혜 미[•°], 허 운 행[*], 김 정 현[**], 박 지 현[**]

Monaural Music-Speech Source Separation Based on Convolutional Neural Network for Background Music Identification in TV Shows

Hyemi Kim[•°], Woon-Haeng Heo[*], Junghyun Kim[**], Jihyun Park[**]

요 약

음악 식별 기술은 비교적 기술적 성숙도가 높지만 이는 입력되는 음원에 손상이나 합성이 없는 경우이고, 방송 오디오와 같이 대사와 음악이 혼재된 상황에서 작은 소리로 혼합된 배경음악을 식별하는 경우 그 성능은 급격히 저하된다. 본 논문에서는 음악-대사 분리 기법을 적용하여 대사가 제거된 음악 신호로부터 배경음악을 식별하고자 한다. 대표적인 컨볼루션 기반 음악 음원 분리 네트워크 구조인 U-Net, Wave-U-Net 및 MMDenseNet 기반 음원 분리를 위한 기존 기법들을 도입하여 음악-대사 분리에 적합하도록 변형한다. 또한 새로운 음악-대사 분리 기법인 DenseNet 구조를 가지는 파형 입력 기반 Wave-DenseNet을 제안한다. 식별기는 랜드마크 기반 오디오 핑거프린트 방식을 적용한다. SDR이 음원 분리 성능 지표로 널리 쓰이고 있으나 분리 후 SDR 값의 성능 순위와 식별율의 성능 순위가 서로 다름으로 보아 분리 후 얻어진 음악 신호로 식별을 하고자 할 때는 적합한 성능 지표가 아님을 확인하였다. 음악-대사 데이터셋으로 음악-대사 분리 후 식별 성능을 비교한 결과 가장 우수한 방법은 Wave-U-Net 분리 기법임을 보였다.

키워드 : 음악 음원 분리, 음악 식별, 음악-대사 음원 분리, 시맨틱 영역 분할, 컨볼루션 신경망
Key Words : music source separation, music identification, music-speech source separation, semantic segmentation, convolutional neural networks

ABSTRACT

Music identification technology has a relatively high technical maturity in the case of the clean music input. However, its performance is drastically reduced when the background music is mixed with speech like TV shows. U-Net, Wave-U-Net, and MMDenseNet and modify them for music-speech separation. Also, we propose Wave-DenseNet which is a waveform input method with DenseNet. A landmark based audio fingerprint method is used for music identification. Although SDR is widely used as a measure of source separation, we confirmed

※ 본 연구는 문화체육관광부 및 한국저작권위원회의 2020년도 저작권기술개발사업의 연구결과로 수행되었습니다. [2018-micro-9500, 음악 및 동영상 모니터링을 위한 지능형 마이크로 식별 기술 개발]
•° First and Corresponding Author : Electronics and Telecommunications Research Institue, miya0404@etri.re.kr, 선임연구원, 정회원
* Chungbuk National University, blue0564@nate.com, 학생회원
** Electronics and Telecommunications Research Institue, bonobono@etri.re.kr; juhyun@etri.re.kr, 정회원
논문번호 : 202001-006-C-RN, Received January 10, 2020; Revised March 6, 2020; Accepted March 6, 2020

it is not a suitable measure for music identification with separated music signal after source separation as the performance rankings of SDR and the identification rate are different. Comparing the background music identification results, we show the best music-speech separation method for the background music identification is the Wave-U-Net based separation method.

Ⅰ. 서 론

오디오 핑거프린팅 기술은 음악으로부터 사람의 지문과 같은 고유의 특성을 추출하여 음악을 식별하는 기술[1-3]이다. 커피 전문점이나 쇼핑몰과 같이 사람들이 웅성거리는 잡음 환경이나 자동차 내부에서 엔진 소리가 들리는 백색 잡음 환경에서는 비교적 우수한 성능을 나타낸다. 그러나 TV에서 방영되는 드라마 등 방송물의 경우 배우가 대사를 하고 있을 때 배경으로 나오는 음악은 배우의 대사보다 소리의 크기가 훨씬 작고 대사가 무척 또렷하여 해당 오디오에서 특징을 추출하면 음악의 고유 특징과 큰 차이가 나게 된다. 이러한 상황에서 음악 식별을 하게 되면 그 성능이 현저히 떨어지게 된다. 작은 소리로 혼합된 배경 음악의 식별 성능을 개선하기 위하여 음악과 대사가 혼합된 입력 신호로부터 바로 특징을 추출하지 않고 먼저 음악과 대사를 분리한 후 대사가 제거된 신호를 이용해 배경 음악을 식별한다면 식별 성능을 높일 수 있을 것이다.

최근 딥러닝 기반 영역 분할 방법이 발전함에 따라 음악에서 각 음원 별 신호를 분리하는 기술 또한 그 성능이 비약적으로 향상되었는데 이와 관련한 대표적인 최신 연구들[4-7]은 skip connection을 가지는 컨볼루션 신경망 형태를 띄고 있다.

Jansson[4]은 의료 영상 분야에서 처음 도입된 기법인 U-Net[8] 구조를 음악의 음원 분리에 적용시켰다. 입력 신호의 진폭 스펙트로그램으로부터 0과 1 사이의 값을 가지는 소프트 마스크를 얻어 음악만의 스펙트로그램을 분할해내고 혼합 신호의 위상 값을 활용해 분리된 음악 신호를 복원하는 방식이다. 이때 U-Net 구조가 가지는 skip connection으로 인해 정교한 영역 분할이 가능한데, 이는 스펙트로그램에서 한 픽셀의 오류로 주파수가 바뀌게 되는 음악 신호의 경우에 적절한 방법이었다. 대용량의 데이터셋을 활용하여 음악에서 가수 목소리의 분리 성능이 높음을 보였고 이후 U-Net을 기반으로 하는 많은 음원 분리 연구가 이어졌다[9-12].

Stoller[5]는 U-Net을 스펙트로그램 입력이 아닌 파형 입력에 적용하였다. 이는 스펙트로그램을 입력으로 사용하는 방식에 비해 다음과 같은 장점이 있다. 스펙트로그램의 경우 진폭값만을 영역 분할에 활용하고 위상 정보는 혼합 신호의 그것을 그대로 적용하는 방식이나, 파형의 경우 위상 정보를 내포하므로 분리 시에 진폭과 위상 정보를 모두 활용할 수 있다. 또한 스펙트로그램보다 더 큰 차원의 1차원 입력이 가능해 네트워크를 더 깊이 구성할 수 있게 된다. 이러한 end-to-end 방식의 음원 분리 방법은 이후 연구[12,13]에서도 활용되었다.

Takahashi는 DenseNet[14]을 기반으로 음원 분리를 위한 멀티스케일 멀티밴드 DenseNet 방식[6]을 제안하였다. DenseNet은 직전 계층의 결과만이 아니라 이전의 모든 계층을 다음 계층의 입력으로 사용하는 방식이다. DenseNet을 음원 분리에 적용하기 위하여 먼저 transposed convolution을 통한 업샘플링 단계를 추가하여 encoder-decoder 형태로 변형한 후 동일 계층끼리 skip connection을 두었다. 학습 시 스펙트로그램의 모든 주파수 대역을 한번에 입력하지 않고 밴드 별로 나눈 후 각각을 DenseNet 기반의 음원 분리 네트워크로 전달한다. 밴드 별 결과를 통합하여 최종 분리된 전체 스펙트로그램을 얻는다. 이후 MMDenseNet에서 나아가 MMDenseNet과 LSTM을 결합한 기법[15]을 제안하기도 하였다.

이러한 음악 음원 분리 최신 기법들을 방송물의 배경 음악 식별 성능 개선을 위한 음악-대사 분리에 적용하기 위해서는 방송물에 등장하는 출연자나 배우의 음성 데이터와 배경 음악을 분리할 수 있는 학습 데이터가 필요하다. 이를 위하여 음악-대사 데이터셋[22]을 활용하였다. 음악-대사 데이터셋은 음악과 대사의 혼합 신호를 합성하여 얻는데 방송 오디오와 유사하게 생성하기 위하여 방송물에서 배경 음악이 전혀 포함되지 않은 대사 구간을 추출하고, 이를 상용 음악과 합성하였다. 합성 시 이들의 신호비를 0dB에서 -30dB 사이의 임의값으로 설정하여 배경 음악에 비해 배우의 목소리가 같거나 더 큰 상황과 유사한 오디오를 생성하였다.

본 논문에서는 음악 음원 분리에 사용되는 컨볼루션 신경망 기반의 부호화-복호화 형태를 갖는 대표적인 네트워크 구조인 U-Net 또는 DenseNet을 활용한

U-Net[4], Wave-U-Net[5] 그리고 MMDenseNet[6] 기법들을 분석하고 이를 방송물에서 음악과 대사를 분리하는 데 적합하도록 변형하여 학습시킨다. 또한 Wave-U- Net이 U-Net을 파형 입력에 적용한 것에 착안하여 본 논문에서는 DenseNet을 파형 입력에 적용한 Wave-DenseNet 기법을 제안하고 제안 기법의 분리 성능을 앞서 서술한 최신 기법들과 함께 비교한다. 성능 비교를 위해 대사와 음악이 다양한 SNR로 합성된 음악-대사 데이터셋에 적용한다. 분리 성능을 측정하는 기준으로 음악과 대사가 혼합된 입력 오디오에서 대사가 제거된 음악 신호를 입력으로 하여 배경 음악을 식별함으로써 분리 기법에 따른 음악-대사 분리 성능을 비교한다. 또한 흔히 분리 성능 측정에 쓰이는 지표[16]에 따른 결과와 비교하여 각 분리 기법이 배경 음악을 식별하는 데 있어 실질적으로 어떤 효과가 있으며 분리 성능 측정 지표에 따른 결과와는 어떤 차이가 있는지를 분석함으로써 최종적으로 배경 음악 식별에 가장 효과적인 분리 기법이 무엇인지 살펴보고자 한다.

Ⅱ. 배경 음악 식별을 위한 음악-대사 분리 기법

음악과 대사를 분리하기 위하여 가수의 목소리를 분리하는 다양한 음원 분리 기법들을 분석한 후 이를 방송 프로그램의 배경 음악 식별을 위한 음악-대사 분리에 적합하도록 수정하여 학습하였다.

가수 목소리 분리를 위한 기존 신경망을 가수 목소리 음원과 반주 음원이 쌍으로 포함된 데이터셋으로 학습한 후 음악 신호의 파형 샘플 또는 진폭 스펙트로그램을 입력하면 가수 목소리와 반주로 분리된 신호를 얻을 수 있다. 이를 음악-대사 분리 학습 신경망에 동일하게 적용할 수 있다. 배경 음악과 방송물의 대사 데이터를 쌍으로 포함한 음악-대사 데이터셋으로 신경망을 학습한 후 혼합 신호의 파형 샘플 또는 진폭

표 1. U-Net 네트워크의 부호화부 구조
Table 1. The structure of the encoder part of U-Net

Layer	Operation	Output size
-	Input	512×128×1
1	conv2D / pooling	256×64×16
2	conv2D / pooling	128×32×32
3	conv2D / pooling	64×16×64
4	conv2D / pooling	32×8×128
5	conv2D / pooling	16×4×256
6	conv2D / pooling	8×2×512

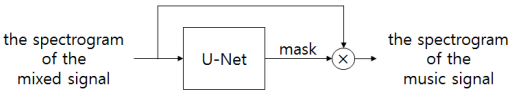

그림 1. U-Net 기반 음악 음원 분리 개념도
Fig. 1. U-Net based music source separation

스펙트로그램을 제공하면 음악만의 파형 샘플 또는 진폭 스펙트로그램을 획득하게 된다. 이렇게 얻어진 분리된 음악을 다시 식별기의 입력으로 제공하면 최종적으로 음악 식별 결과를 얻을 수 있다.

2.1 U-Net 기반 음원 분리 방법

영상 인식 분야에서 원하는 객체를 찾기 위하여 딥러닝 기반 시맨틱 영역 분할 방법을 사용하면 해당 객체의 픽셀이 포함된 영역 분할 마스크를 얻을 수 있다[17]. U-Net[4]에서는 이러한 시맨틱 영역 분할 방식을 음원의 스펙트로그램 영상에 적용하였다.

1차원 오디오 신호를 영상과 유사한 형태인 2차원 정보로 변환하기 위하여 STFT(Short-Time Fourier Transform)을 적용한 후 진폭 값만을 취하여 2차원의 스펙트로그램 영상을 획득한다. 그림 1과 같이 음악과 대사가 혼합된 신호로부터 얻은 스펙트로그램 영상에서 음악을 분할하는 소프트 마스크를 얻은 후, 혼합 신호의 스펙트로그램과 마스크의 곱을 통해 음악 신호만의 스펙트로그램을 얻는다.

음악 신호는 한 픽셀의 오류로 주파수가 변할 수 있기 때문에 영상 신호에 비해 매우 높은 정확도가 요구된다. 이러한 고정밀의 소프트 마스크를 얻기 위하여 skip connection을 가지는 U-Net[4] 방법을 적용하였다. 이 방법은 일반 영상에 비해 상대적으로 높은 정밀도가 요구되는 의료 영상 분야에서 처음 도입되어 그 성능이 검증되었다[8].

U-Net 기반 음악-대사 분리 네트워크는 U-Net[4]의 네트워크 구조와 동일하게 구성하였다. 네트워크 구조는 표 1과 같이 모든 계층이 컨볼루션 신경망으로 구성되어 있다. 부호화부와 복호화부는 각각 6개의 계층으로 구성된다. 동일한 계층의 부호화부와 복호화부 간에 skip connection을 구성함으로써 음원 분리를 더욱 정교하게 한다. 네트워크를 거쳐 얻어진 소프트 마스크를 초기 입력인 혼합 신호 스펙트로그램과 곱하여 최종적으로 분리된 신호의 스펙트로그램을 획득한다.

음악-대사 분리 네트워크의 구체적인 학습 과정은 다음과 같다. 먼저 음악과 대사 각각의 음원 파일을 준비하고 이를 원하는 신호비를 적용하여 혼합한 학습 데이터셋을 준비한다. 입력으로 주어진 혼합 신호

의 스펙트로그램으로부터 음악 신호의 스펙트로그램을 추출하는 네트워크를 학습하기 위하여 먼저 다음과 같이 학습 데이터셋으로부터 음악 신호의 소프트 마스크를 계산한다. 음악 신호의 스펙트로그램 내 각 bin 별 값을 혼합 신호 스펙트로그램의 동일한 bin의 값으로 나누면 전체 bin에 대해 0과 1 사이의 값을 가지는 신호 분리 마스크를 얻을 수 있다. 이를 학습할 네트워크의 출력값으로 제공한다. 이때 네트워크 학습에 쓰일 손실함수는 다음과 같다.

$$L(X, Y;\Theta) = \| f(X,\Theta)\odot X - Y \|_{1,1} \quad (1)$$

위 식에서 X는 음악과 대사가 혼합된 스펙트로그램, Y는 대사의 스펙트로그램, $f(X,\Theta)$는 네트워크 파라미터를 이용하여 혼합된 신호로부터 음성만을 추출해내는 소프트 마스크이며, $L_{1,1}$ 행렬 norm은 각 원소의 절대값의 합으로 계산된다.

U-Net[4]에서 가수 목소리 분리를 위하여 입력 오디오의 샘플링 주파수는 8,192Hz로 하고 FFT 수행 시 윈도우 크기를 1024, 입력 스펙트로그램의 프레임 길이는 128, hop 크기는 768로 하였다. 음악과 대사 분리를 위한 방송 오디오 입력은 고품질의 오디오를 사용하므로 본 논문에서는 샘플링 주파수를 16kHz로 설정한 후 입력 프레임 길이와 중첩 정도에 따른 식별율 변화를 살펴보고 음악-대사 분리 성능을 더욱 개선할 수 있는 변수값을 찾는 실험을 하였다.

2.2 Wave-U-Net 기반 음원 분리 방법

2.1절에서 서술한 U-Net 기반 음원 분리 방식은 입력 음원으로부터 STFT를 수행한 후 얻어진 스펙트로그램의 진폭과 위상 중 진폭 정보만을 활용한다. 이러한 방식은 다음과 같은 한계점이 있는데 먼저 푸리에 변환 시 윈도우 크기와 중첩 값에 결과가 의존적이라는 점이다. 또한 분리 후 오디오 파일로 복원 시 혼합 신호의 위상 정보를 그대로 사용하게 된다. 그리고 분리 과정에서 위상 정보를 완전히 배제하여 성능을 제한시킨다. 이러한 한계를 극복하기 위하여 Stoller가 제안한 위상 정보를 직접적으로 사용할 수 있는 파형 정보를 이용하는 방식인 Wave-U-Net[5]은 주어진 입력 정보를 최대한 활용한다는 측면에서 더 효과적이다.

네트워크 구조는 그림 2과 같다. U-Net[4]과 동일한 부호화부-복호화부 구조를 띄며 U-Net[4]에서 입력으로 사용된 스펙트로그램을 다루기 위한 2차원 입력 대신 파형의 1차원 정보를 다루는 형태로 변형되었다. 입력 샘플의 수가 스펙트로그램에 비해 커짐으로 인해 pooling 횟수를 늘릴 수 있어 계층 수를 12로 설정 가능하여 더욱 깊은 네트워크를 형성할 수 있게 된다.

기존 방식에서 주로 사용되는 업샘플링 방식은 strided transposed convolution 방법이다. 영상에서 픽셀 사이에 0을 추가하여 업샘플링을 적용한 후 컨볼루션을 적용하는 방식인 strided transposed convolution을 그대로 파형 입력에 대해 적용한다면 샘플 사이에 0의 값을 채워줌으로 인해 값의 큰 변화에 따른 고주파 성분의 artifact가 생성되기 쉽다. Stoller는 이러한 artifact 생성을 방지하기 위하여 strided transposed convolution 방식 대신 선형보간법을 채택하였다. 0 대신에 두 샘플의 선형보간 값을 두 샘플 사이에 추가하여 artifact를 줄이면서 업샘플링을

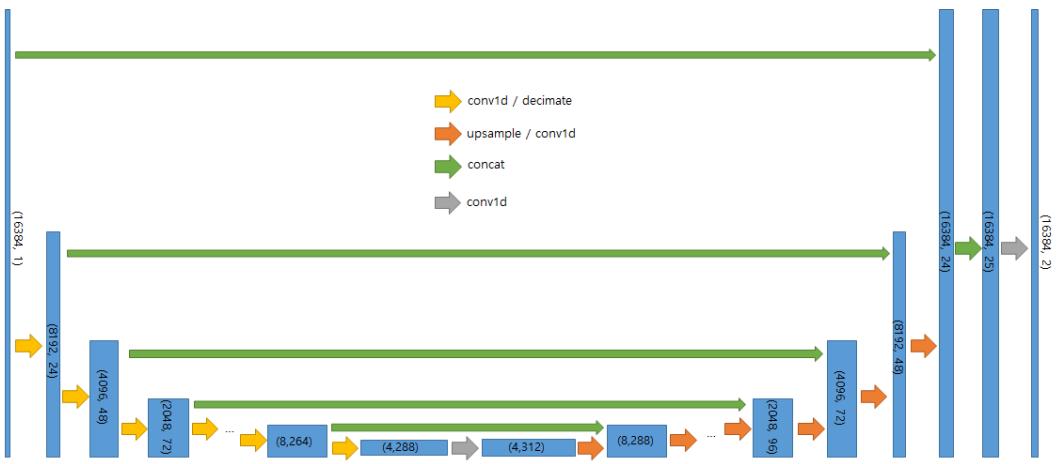

그림 2. 음원 분리를 위한 Wave-U-Net 기본 구조
Fig. 2. The structure of Wave-U-Net for music source separation

수행하였다.

Wave-U-Net[5]에서는 이러한 문제가 다운샘플링 과정에서 컨볼루션 시 샘플의 경계에서도 동일하게 발생할 것이라 보고 zero-padding을 적용하지 않았다. 하지만 이는 출력 신호가 입력 신호의 길이보다 훨씬 짧아져 신호 분리 수행 시간이 급격히 증가하는 문제가 있다. 본 실험에서는 컨볼루션 시 zero-padding을 적용하고 업샘플링 시 선형 보간법을 적용하여 입력 샘플과 동일한 길이의 샘플을 출력으로 얻도록 구조를 변경하였다. 즉, Wave-U-Net[5]의 다양한 모델 중 기본 구조인 M1 모델을 적용하였다. Wave-U-Net[5]의 실험 결과를 보더라도 스테레오 입력을 사용한 M4 모델에서 분리 성능이 큰 폭으로 향상되나 모노채널을 적용한 그 외 모델에서는 기본 모델에 비해 성능 향상이 크지 않음을 확인할 수 있다.

2.3 MMDenseNet 기반 음원 분리 방법

Takahashi는 DenseNet을 기반으로 하는 음원 분리를 위한 멀티스케일 멀티밴드 DenseNet 방식[6]을 제안하였다. DenseNet[14]은 당시 영상 분류 분야에서 최고 성능을 내던 ResNet[18]의 문제점을 해결하기 위하여 제안되었다. 네트워크를 학습할 때 ResNet은 이전 계층 x_{l-1}의 결과를 다음 계층 x_l의 입력으로 사용한다. 여기에 skip-connection을 추가하여 이전 계층의 결과를 더함으로써 최종적으로 다음 계층의 출력을 얻는다. 즉, $H_l(\bullet)$ 이 l번째 비선형 변환을 나타낼 때 그의 출력은 다음과 같다.

$$x_l = H_l(x_{l-1}) + x_{l-1} \tag{2}$$

이러한 skip-connection을 통해 그레디언트를 얻는데 도움을 주지만 이전 정보를 현재 계층의 결과와 더함으로써 정보의 흐름을 방해하는 단점이 있다[14].

이를 개선하기 위하여 DenseNet에서는 직전 계층의 결과만이 아니라 이전의 모든 계층을 다음 계층의 입력으로 사용한다. 또한 ResNet에서 skip-connection을 더하여 사용한 것과 달리 DenseNet에서는 이전 정보를 직접 concatenation하여 재사용한다. 즉, l번째 계층은 모든 이전 계층의 특징맵 $x_0, \cdots, x_{l-1}$으로부터 다음과 같이 얻을 수 있다.

$$x_l = H_l([x_0, x_1, \cdots, x_{l-1}]) \tag{3}$$

이때 $H_l(\bullet)$는 배치 정규화와 ReLU activation 함수, 그리고 3×3 컨볼루션으로 구성된다.

다운샘플링 수행을 위하여 이전의 모든 계층과 조밀하게 연결된 l개의 계층을 묶어주고 이를 dense block이라 명명한다. 이러한 dense block들 사이에 transition layer를 두어 배치 정규화와 1×1 컨볼루션, 그리고 2×2 average pooling을 수행한다.

$H_l(\bullet)$이 k개의 특징맵을 생성한다고 할 때 입력의 채널 개수를 k_0라고 하면 l계층에서의 특징맵은 $k_0 + k\times(l-1)$개가 된다. k를 growth rate라고 하였고, 상대적으로 작은 값의 k만으로도 당시 최신 기법 수준의 결과를 얻을 수 있음을 보여주었다.

Takahashi는 이러한 DenseNet을 기반으로 영상이 아닌 오디오에 적용하여 음원을 분리하는 방법을 제안하였다. DenseNet을 부호화부-복호화부 형태로 변형하기 위하여 transposed convolution을 통한 업샘플링 단계를 추가하고 U-Net[4]과 유사하게 동일 계층의 denseblock의 결과를 skip-connection을 통해 concatenation한다. 이렇게 변형된 DenseNet을 입력 스펙트로그램 전체에 대해 한번만 적용하는 것이 아니라 스펙트로그램을 N개의 밴드로 나눈 후, 각 밴드 및 전체에 대해 DenseNet을 각각 적용하고 결과를 통합하는 방식을 제안하였다. 논문에서 설정한 입력 음원의 샘플링 주파수는 44.1kHz, STFT 수행 시 윈도우 크기는 2,048로 50%씩 겹쳐 나가면서 STFT를 수행하였다.

기존 가수 목소리 분리 기법들의 성능 비교 실험 결과를 살펴보면[7] MMDenseNet이 U-Net보다 우수함을 확인하였다. 하지만 본 실험에서는 분리에서 그치는 것이 아니라 네트워크의 구조에 따른 음악 분리 후 배경음악의 식별 성능 개선 정도를 비교하고자 하므로 입력되는 스펙트로그램을 생성할 때 쓰이는 파라미터들을 서로 유사한 수준으로 설정하였다. 입력 음원의 샘플링 주파수는 16kHz, STFT 윈도우 크기는 1024, 중첩 정도는 75%, 입력 프레임 개수를 512로 변경하여 실험하였다. 멀티밴드 기법을 적용하기 위하여 N을 2로 하고 스펙트로그램을 저주파수 대역과 고주파수 대역으로 나누었다. MMDenseNet[6]은 각 음원 별로 네트워크를 각각 학습한 후 분리하였으나 본 실험에서는 U-Net 및 Wave-U-Net과 동일하게 음악과 대사 둘로 동시에 분리 가능하도록 수정하였다. 혼합 신호에서 분리된 음악 신호의 스펙트로그램을 얻을 수 있는 소프트 마스크 값을 활용하여 대사 신호의 스펙트로그램을 구함으로써 음악 신호와 대사 신호를 동시에 분리하고 복원하였다.

Ⅲ. 제안 방법

2.3절에서 서술한 MMDenseNet은 DenseNet을 스펙트로그램 입력에 대하여 적용한 기법이다. 본 논문에서는 Wave-U-Net이 U-Net의 네트워크 구조를 파형 입력에 대하여 적용한 것에 착안하여 DenseNet을 파형 입력에 대하여 적용한 새로운 Wave-DenseNet 기법을 제안하고 음원 분리를 적용한 후 얻어진 음악 신호에 대해 음악 식별 성능 비교 실험을 수행하였다.

입력 신호의 샘플 크기는 Wave-U-Net과 동일한 16,384로 하고 DenseNet의 growth rate k와 계층 수 L은 각각 12로 하였다. 부호화부에서 쓰일 변이층(transition layer)는 5개, 1×1 컨볼루션 시 적용될 압축률(compression rate)은 0.2로 하였다. 컨볼루션의 필터 크기는 15로 하였다. 복호화부에서는 transposed convolution을 적용하여 샘플의 길이를 초기 입력과 동일한 크기로 변환하였다. Wave-U-Net에서는 복호화부의 컨볼루션 필터 크기를 5로 작게 하였으나 본 논문에서 제안한 기법은 부호화부와 복호화부의 컨볼루션 크기를 15로 동일하게 하였다. 네트워크의 흐름도 및 상세 구조는 각각 그림 3과 표 2에 나타내었다.

표 2. Wave-DenseNet 네트워크 세부 작업 수행표
Table 2. The detailed operations of Wave-DenseNet

Block	Operation
Initial conv	$[15\times1$ conv$]$ (k)
DenseBlock and Transition Down for $i=1,\cdots,s$	$[15\times1$ conv$]$ (k,L) $[1\times1$ conv$]$ 2×1 average pool
DenseBlock for $i=s+1$	$[15\times1$ conv$]$ (k,L)
Transition Up and DenseBlock for $i=s,\cdots,1$	$[1\times1$ conv$]$ $[15\times1$ transconv$]$ concat(denseblock i) $[15\times1$ conv$]$ (k,L)
Final conv	$[1\times1$ conv$]$ $[15\times1$ conv$]$ $(4,2)$ $[1\times1$ conv$]$ $[15\times1$ conv$]$ $(2,1)$

Ⅳ. 랜드마크 기반 오디오 식별 방법

음원 분리 기술을 적용하여 혼합 음원으로부터 분리된 음악 신호에 대해 식별 성능이 개선되었는지를 확인하기 위한 식별 방법으로 랜드마크 기반의 오디오 핑거프린팅 방법[1]을 적용한다. 음악 식별에 있어 좋은 성능을 보이는 다양한 식별 기법들이 존재하나[2][3] 본 실험은 음악-대사 분리 기법에 따른 식별 성능 개선 정도의 경향성을 확인하고자 하므로 기본적인 랜드마크 기반 식별 기법[1]을 활용한다.

음악을 식별함에 있어 주변 환경에 소음이 존재하더라도 소리의 온셋 정보는 보존될 것이라는 전제하에, 신호의 온셋 정보, 즉 스펙트로그램의 피크 값들을 추출한다. 추출된 피크들을 시간과 주파수 값을 살펴 가까운 피크들끼리 쌍을 짓고 해당 음악의 해시 정보로 저장한다. 쿼리 음악을 식별하기 위하여 앞서 서술한 방법으로 랜드마크 정보를 획득한 후 데이터셋에 저장된 정보와 비교한다. 동일한 랜드마크 정보의 양이 일정 수준 이상이 되면 해당 곡명을 식별 결과로 출력한다.

Ⅴ. 실 험

2장에서 언급한 다양한 음원 분리 기법들[4-6]은 가수의 목소리를 분리하거나 음악을 이루는 각 악기별 소리를 분리하는 데이터셋[19-21]을 활용하여 그 성능을 보였다. 그러나 이들이 모두 동일한 데이터셋에서 비교되지 않아 그 성능을 서로 정확히 비교하기가 어려웠다.

학습 데이터의 양은 성능에 크게 영향을 미친다. 예를 들어 U-Net 기반 가수 목소리 분리 기법이 우수한 성능을 가짐을 보였으나[4] 이는 상업 음악 2만 곡의 원곡과 반주 쌍으로부터 학습한 결과이다. 다른 논문들이 활용한 데이터셋의 규모에 비해 지나치게 큰 규

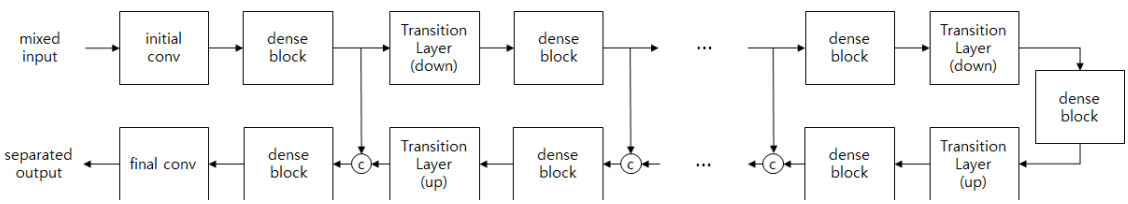

그림 3. 음원 분리를 위한 Wave-DenseNet 네트워크 흐름도
Fig. 3. The flow of Wave-DenseNet for music source separation

모로, 보컬 분리에서 흔히 쓰이는 데이터셋 중 하나인 MIR-1K 데이터셋으로 학습한 결과[7]를 보면 그 성능이 확연히 떨어짐을 알 수 있다.

또한 현재 음원 분리의 결과 비교를 위한 지표로써 많은 연구들에서 SDR이 주로 다루어지고[5-7] 있는데 SDR이 원음을 얼마나 잘 표현하는지에 대한 지표로 널리 쓰이고는 있지만 SDR이 음악 분리 후 식별 성능 개선 여부를 판별하는 데에도 역시 적합한 지표인지에 대해서는 확인된 바가 없다.

이에 따라 방송물의 배경 음악 식별을 위한 사전 분리 방법의 성능 비교 실험을 위하여 음악-대사 분리 데이터셋[22]을 활용하여 동일한 학습 데이터와 검증 데이터로 음원 분리가 식별 성능에 미치는 영향을 분석하고, 그 결과를 기존 분리 성능 측정 방법인 SDR 값과 비교해보고자 한다.

5.1 음악-대사 신호 분리 데이터셋

음악-대사 신호 분리 성능 비교를 위한 테스트용 데이터로 실제 방송물을 활용한다면 본 실험의 목적에 가장 잘 부합하겠으나 정확한 배경 음악 큐시트를 다량 확보하는 데는 큰 어려움이 따른다. 따라서 본 논문에서는 방송물의 오디오 상황과 유사한 데이터셋을 직접 생성하여 이를 학습 및 검증에 사용하였다.

음악-대사 분리 데이터셋은 학습과 검증용으로 각각 12초 길이의 1,823개 음악과 동일 길이의 대사, 그리고 이들의 혼합 신호로 이루어져 있다. 다양한 장르의 방송 프로그램으로부터 100시간의 오디오를 수집 및 추출한 후 배경 음악이 사용되지 않고 대사만 존재하는 구간을 태깅하여 대사 데이터셋을 생성하였다.

학습을 위한 혼합 신호는 SNR 값을 -30dB부터 0dB까지의 임의의 값으로 음악 신호와 대사 신호를 혼합하여 만들었다. 다음으로 검증용 혼합 신호는 음악 대 대사 소리 크기에 다른 식별율 변화를 확인하기 위하여, 음악과 대사가 동일한 크기로 혼합되어 SNR이 0dB인 경우와 음악이 배경 음악으로 사용되는 수준인 -10dB 두 경우로 구성하였다. 동일한 방식으로 샘플링 주파수를 다르게 하여 16kHz 데이터셋과 22,050Hz 데이터셋을 생성하였다.

식별을 위한 음악 데이터셋은 음악-대사 분리 데이터셋에서 사용된 3,646곡을 포함한 다양한 장르의 국내외 대중가요 9,118곡으로 구성되어 있다. 대부분 대사와 혼동될 수 있는 여지가 큰 보컬이 포함되어 있고 식별을 위한 음악 데이터셋 내 9,118곡의 전 구간에서 음악 식별을 위한 랜드마크 정보를 추출하여 데이터베이스로 저장하였다.

5.2 프레임 길이에 따른 실험 결과

Jansson은 U-Net[4]에서 음원의 샘플링 주파수를 8,192Hz로 하고 푸리에 변환 시 윈도우 크기를 1,024로 25% 중첩하였다. 이를 Wave-U-Net[5]과 유사한 환경으로 설정하기 위하여 샘플링 주파수는 16kHz로, 윈도우 크기는 1024로 변경하고 중첩 시 hop 크기는 373으로 고정한 후 학습 시 입력 스펙트로그램의 프레임 수에 따른 식별 성능 변화를 관찰한 결과 표 3과 같은 결과를 얻었다. 12초 길이의 입력 음원에 대해 음원 분리를 적용하는 경우 학습 네트워크의 입력이 각각 128, 256 프레임인 경우 입력 오디오를 해당 프레임 크기로 잘라 분리 마스크를 얻은 후 이들을 다시 시간 축으로 연결하여 동일한 길이의 분리 음원을 획득한다. U-Net 논문에서 제안한 128 프레임의 경우와 비교하여 프레임 수가 512인 경우 25.6%에서 38.7%로 큰 폭의 성능 향상을 확인하였다. 즉, 동일한 분리 네트워크 모델을 사용하더라도 학습 시 네트워크의 입력 단위 길이를 길게 하면 성능이 향상됨을 확인하였다.

5.3 Hop 크기에 따른 실험 결과

22,050Hz의 샘플링 주파수에 대해 학습 단위의 프레임 수를 512로, 푸리에 변환 윈도우 크기는 1024로 고정한 후 hop 크기를 윈도우 크기의 2/8, 3/8, 4/8에

표 3. U-Net 기반 대사-음악 분리 적용 후 학습 네트워크 입력 프레임 수에 따른 식별율 비교 (단위 : %)
Table 3. The music identification rate according to the number of frames after U-Net based music-speech source separation (unit : %)

Number of frames (duration)	Recognition rate	
	0dB	-10dB
128 (3.02s)	25.6	7.6
256 (6.00s)	28.7	13.2
512 (11.98s)	**38.7**	**18.6**

표 4. U-Net 기반 대사-음악 분리 적용 후 hop 크기에 따른 식별율 비교 (단위 : %)
Table 4. The music identification rate acoording to the hop size after U-Net based music-speech source separation (unit : %)

hop size	Recognition rate	
	0dB	-10dB
256	31.9	**15.1**
384	**32.3**	13.3
512	27.9	11.0

해당하는 256, 384, 512로 변화시키면서 분리 후 식별율을 측정하였다. 동일 iteration을 수행한 결과 표 4에서 보면 Hop 크기가 512인 경우, 즉 윈도우를 50% 겹쳐 나가면서 푸리에 변환을 수행하는 것보다 50% 보다 많이 겹치는 경우 식별율이 더 높음을 확인하였다. 또한 hop 크기가 그보다 작은 경우에는 0dB인 경우 hop 크기가 384일 때, -10dB인 경우 hop 크기가 256일 때 가장 식별율이 좋으나 그 차이는 미미하다. 이로 보아 hop의 크기가 작을수록 항상 성능이 좋아지는 것은 아니며, FFT 윈도우의 50% 이상을 겹쳐나가면서 동일 길이의 음원에 대해 계산 효율이 더 좋은 hop 크기를 선택하는 것이 성능 저하는 크게 없는 대신 실행 속도에서 이득을 얻을 수 있을 것으로 보인다.

5.4 음악-대사 음원 분리 성능 측정 결과

음악-대사 분리 성능을 측정하기 위하여 음원 분리 분야에서 널리 쓰이는 BSS Eval[16]을 활용하여 분리 성능 지표인 SDR (signal-to-distortion ratio)과 SIR(source-to-interference ratio), 그리고 SAR (source-to-artifacts ratio)를 측정하였다. 측정 시 mir-eval 툴박스[23]를 사용하였다.

표 5에서 음악의 분리 성능을 살펴보면 SDR의 경우 MMDenseNet이 가장 우수한 분리 성능을 보이고 그 뒤로 Wave-U-Net, U-Net, Wave-DenseNet 순으로 성능이 우수하다. SIR의 경우 Wave-U-Net이 특히 우수한 성능을 보이고 Wave-DenseNet이 다음으로 우수한 결과를 보인다. 대사의 분리 성능은 SDR 값의 경우 0dB에서는 MMDenseNet이, -10dB에서는 Wave-U-Net이 가장 좋은 성능을 보였다. SIR 값은 MMDenseNet이 가장 좋은 성능을 보였고 SAR 값은 Wave-U-Net이 가장 우수하였다.

이러한 결과는 다음 절에서 보일 음악 식별 성능 결과와 약간의 차이를 보인다. 특히 SDR은 기존의 많은 음원 분리 기법들[5-7]이 분리 성능 지표로 사용하고 있다. 다음 절에서 우리는 분리된 음원의 식별 성능을 직접적으로 측정함으로써 SDR 값으로 측정한 기법별 분리 성능 순위와 음악 식별 성능간의 차이를 보이고자 한다.

5.5 분리된 음악 식별 실험 결과

음악-대사 분리 후 분리된 음악에 대한 식별율을 측정하였다. 식별 방식은 랜드마크 기반의 오디오 핑거프린팅 기법[1]을 적용하였다. 표 6은 혼합 음원으로부터 각각의 분리 기법을 적용한 후 분리된 음악을 입력으로 하여 음악 식별을 수행한 결과이다.

먼저 입력 음원을 11,025Hz로 다운샘플링 하고 스펙트로그램 생성 시 푸리에 변환 윈도우 크기는 512, hop 크기는 256으로 한 경우(p1)의 식별 결과를 살펴보면, 음악과 대사가 동일한 소리 크기로 혼합된 0dB의 경우 분리 수행 전 혼합 음원(mix)의 식별율은 40.5%였으나 U-Net을 적용하여 분리한 후 음악을 식별한 결과 39.3%로 분리 후 음악 식별 성능이 1.2% 가량 저하되었다. 그러나 U-Net을 제외한 Wave-U-Net, MMDenseNet, Wave-DenseNet 기반 기법들은 분리 후 성능 개선이 이루어졌고, 특히 파형 입력을 기반으로 하는 Wave-U-Net과 Wave-DenseNet은 각각 60.7%와 56.6%로 성능이 큰 폭으로 개선됨을 확인하였다.

다음으로 대사가 크게 혼합되고 음악이 배경 음악으로 작게 혼합된 -10dB의 경우 분리 전 식별 성능이 3.8%로 성능이 매우 나빴으나 분리 기법을 적용한 후 식별 성능은 약 5~9배가량 향상되었다. 파형 기반의 두 기법은 약 7배 이상의 성능 향상을 보였다. 음악-대사 음원 분리를 적용하면 배경 음악으로 작게 삽입된 음악의 식별 성능을 향상시키는 데 특히 도움이 됨을 실험적으로 확인하였다.

기법 별로 보면 U-Net의 식별 성능 개선 정도가 가장 적었고 다음으로는 같은 스펙트로그램 기반 분리 기법인 MMDenseNet의 성능 개선 정도가 낮았다. 가장 우수한 기법은 Wave-U-Net을 적용한 경우였고 다

표 5. 기법별 음원 분리 성능 측정
Table 5. The evaluation of source separation methods

SNR	0dB								-10dB							
method	U-Net		Wave-U-Net		MMDenseNet		Wave-DenseNet		U-Net		Wave-U-Net		MMDenseNet		Wave-DenseNet	
source	Mu	Sp	Mu	Sp	Mu	Sp	Mu	Sp	Mu	Sp	Mu	Sp	Mu	Sp	Mu	Sp
SDR	6.26	5.86	6.67	6.48	**6.91**	**6.74**	5.19	5.11	3.18	14.48	3.48	**14.68**	**3.61**	14.52	2.17	13.79
SIR	12.28	10.07	**15.26**	11.1	12.26	**12.49**	13.37	8.73	10.11	19.09	**14.53**	19.37	9.92	**20.49**	12.49	17.72
SAR	8.11	9.08	7.65	**9.27**	**9.03**	9.00	6.38	8.93	4.98	16.81	4.29	**16.86**	**5.54**	16.13	3.23	16.59

음으로는 Wave-DenseNet이 성능 개선 정도가 높았다.

다음으로 식별 파라미터 중 샘플링 주파수는 16kHz, 푸리에 변환 윈도우 크기는 1024, hop 크기는 373으로 변경하여 식별율을 측정한 결과(p2)를 살펴본다. 이 경우 분리 적용 후 식별율이 분리 전에 비해 모두 개선되었고 성능 개선 정도는 앞서 설정한 식별기 파라미터의 경우와 동일하게 파형 기반 방식의 성능 개선이 가장 컸고, 특히 Wave-U-Net 기반 방식이 가장 우수하였다.

파라미터를 p1에서 p2로 변경한 경우 약 9천여 곡에 대한 오디오 핑거프린트 정보를 저장한 데이터베이스의 크기는 190MB에서 317MB로 약 1.7배의 저장 공간을 필요로 하였고, 약 6시간 분량의 음악 식별 시간은 약 63초에서 94초로 약 1.5배 길어졌으나 식별 성능은 현저히 향상됨을 확인하였다. 식별 시간이 길어졌다고 하지만 1시간 분량의 입력이 15초 이내로 식별되는 점, 그리고 저장 공간은 손쉽게 늘일 수 있다는 점을 감안하면 식별 파라미터는 개선된 파라미터 값을 설정하여 식별하는 것이 저장 공간이나 식별 시간에 대한 큰 부담 없이 우수한 성능을 얻는 데 유리함을 확인하였다.

네트워크의 파라미터 개수를 살펴보면 Wave-U-Net의 파라미터 개수는 약 천만 개이나 Wave-DenseNet은 약 300만 개로 Wave-U-Net에 비해 학습 모델의 저장 용량을 약 3분의 1 가량 줄일 수 있었다.

앞서 5.4절에서 살펴본 것과 같이 분리된 음악 신호로 식별할 경우 음원 분야에서 대표적으로 쓰이는 음원 분리 성능 지표인 SDR의 성능 순위와는 차이가 있음을 확인하였다. 분리된 음악 신호의 SIR 값을 살펴보았을 때 가장 우수한 결과를 보이는 기법은 Wave-U-Net으로 음악 식별 결과와 동일하였다.

표 6. 음원 분리 기법에 따른 음악 식별율 비교 (단위 : %)
Table 6. The music identification rate according to music-speech separation methods (unit : %)

fingerprint parameter	p1		p2	
SNR	0dB	-10dB	0dB	-10dB
mix	40.5	3.8	64.2	13.4
U-Net[4]	39.3	19.4	69.1	43.0
Wave-U-Net[5]	**60.7**	**35.1**	**82.0**	**59.3**
MMDenseNet[6]	48.1	24.0	76.6	51.7
Wave-DenseNet	56.6	30.0	80.9	55.8

그러나 SDR은 MMDenseNet이 가장 우수한 성능을 나타냄에 반해 음악 식별율 측정 실험에서는 Wave-U-Net이 가장 우수함을 확인하였다. MMDenseNet과 Wave-DenseNet을 비교한 결과 또한 SDR은 MMDenseNet이 우수하나 음악 식별율 측정 실험에서는 Wave-DenseNet이 상대적으로 우수한 성능을 나타내는 것을 확인하였다.

5.6 원 음악으로부터 음악-대사 분리 후 식별 실험

우수한 성능의 음악-대사 분리 기술이라면 대사가 없이 음악만 입력되는 경우에도 음악 전체를 최대한 분리해내어 음악 신호의 손상을 최소화하여야 한다. 이를 확인하기 위하여 음악과 대사를 혼합할 때 쓰였던 음악 원곡에 대해 음악-대사 분리 기법을 적용한 후 분리된 음악 결과로 음악을 식별하였다.

원곡을 입력한 후 다양한 음악-대사 분리를 수행한 결과 표 7과 같은 결과를 얻었다. 식별 파라미터 p2를 적용한 분리 전 원곡 음악의 식별율은 98.0%였지만 분리 후 모든 방법에서 식별율이 저하되었다. U-Net을 적용한 경우 식별율이 가장 큰 폭으로 떨어졌고, 다음으로는 MMDenseNet이었다. 가장 소폭으로 식별율이 저하된 방법은 Wave-DenseNet을 적용한 경우로 원 음악에 음원 분리를 적용한 후의 식별율인 95.3%와 98.0%에서 각각 9.5%와 2.4% 떨어진 85.8%와 95.6%의 식별율을 얻었다. Wave-U-Net을 적용한 경우는 각각 84.9%와 95.3%로 Wave-DenseNet과 근소한 차이를 보였다. 이로 보아 Wave-DenseNet이 원음을 가장 적게 훼손하면서 음악과 대사를 분리하는 방법임을 확인하였다.

표 7. 원 음악 입력에 대한 음원 분리 적용 후 식별율 비교 (단위 : %)
Table 7. Music recognition rate from original music according to music-speech separation methods (unit : %)

Fingerprint parameter	p1	p2
original music	95.3	98.0
U-Net[4]	56.0	86.1
Wave-U-Net[5]	84.9	95.3
MMDenseNet[6]	77.2	93.2
Wave-DenseNet	**85.8**	**95.6**

Ⅵ. 결 론

지금까지 본 논문에서는 방송물에서 배우의 대사가 존재하는 상황에서 배경 음악으로 작게 삽입된 음악

을 식별하기 위하여 최근 음악 음원 분리 분야에서 활용되는 대표적인 기법들에 관해 서술하였고 이를 음악-대사 분리에 적합하도록 변형하였다. 또한 DenseNet 기반의 음원 분리 기법을 파형 입력에 적용한 Wave-DenseNet을 제안하였다. 분리 후 얻은 음악 신호에 대해 BSS Eval 분리 성능 측정 방법과 오디오 핑거프린팅 기반 음악 식별율 측정 방법을 적용하였고 그 결과를 비교하였다.

SDR이 음원 분리 성능 지표로 널리 쓰이고 있으나 분리 후 얻어진 음악 신호로 식별을 하고자 할 때는 적합한 성능 지표가 아님을 확인하였다. 분리 후 SDR을 살펴본 결과 MMDenseNet을 적용한 경우가 가장 우수한 성능을 나타내었으나, 혼합된 음원에서 음악을 분리한 후 식별할 때는 Wave-U-Net을 적용한 경우의 식별율이 가장 높았다. 배우의 대사 없이 음악만 존재하는 경우 분리 기법을 적용하여 성능을 비교한 경우 Wave-U-Net의 결과도 우수한 편이었으나 상대적으로 Wave-DenseNet의 성능이 가장 우수함을 확인하였다.

References

[1] A. Wang, "An industrial-strength audio search algorithm," in *Proc. Int. Conf. Music Info. Retrieval,* pp. 7-13, 2003.

[2] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system," in *Proc. Int. Soc. Music Info. Retrieval Conf.*, pp. 107-115, 2002.

[3] B. Coover and J. Han, "A power mask based audio fingerprint," in *Proc. IEEE Int. Conf. Acoustics, Speech and Sign. Process.*, pp. 1394-1398, Florence, Italy, May 2014.

[4] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-Net convolutional networks," in *Proc. Int. Soc. Music Info. Retrieval Conf.*, pp. 745-751, Suzhou, China, 2017.

[5] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A multi-scale neural network for end-to-end audio source separation," in *Proc. 19th Int. Soc. Music Info. Retrieval Conf.*, pp. 334-340, Paris, France, 2018.

[6] N. Takahashi and Y. Mitsufuji, "Multi-scale multi-band DenseNets for audio source separation," in *Proc. IEEE Workshop on Appl. Sign. Process. Audio and Acoust.*, pp. 21-25, 2017.

[7] S. Park, T. Kim, K. Lee, and N. Kwak, "Music source separation using stacked hourglass networks," in *Proc. Int.Soc. Music Info. Retrieval Conf.,* pp. 289-296, Paris, France, Sep. 2018.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Computing and Computer-Assisted Intervention, Springer*, pp. 234-241, 2015.

[9] D. Stoller, S. Ewert, and S. Dixon, "Adversarial semi-supervised audio source separation applied to singing voice extraction," in *Proc. IEEE Int. Conf. Acoust., Speech and Sign. Process.*, pp. 2391-2395, 2018.

[10] L. Prétet, R. Hennequin, J. Royo-Letelier, and A. Vaglio, "Singing voice separation: A study on training data," in *Proc. IEEE Int. Conf. Acoustics, Speech and Sign. Process.*, pp. 506-510, Brighton, UK, May 2019.

[11] A. Jansson, R. M. Bittner, S. Ewert, and T. Weyde, "Joint singing voice separation and F0 estimation with deep U-Net architectures," in *Proc. 27th Eur. Sign. Process. Conf.*, A Coruna, Spain, Sep. 2019.

[12] A. Cohen-Hadria, A. Roebel, and G. Peeters, "Improving singing voice separation using deep U-net and Wave-U-Net with data augmentation," in *Proc. 27th Eur. Sign. Process. Conf.*, 2019.

[13] O. Slizovskaia, L, Kim, G, Haro, and E. Gomez, "End-To-End sound source separation conditioned on instrument labels," in *Proc. IEEE Int. Conf. Acoustics, Speech and Sign. Process.*, pp. 306-310, Brighton, UK, May 2019.

[14] G. Huang, Z. Liu, L. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 4700-4708, 2017.

[15] N. Takahashi, N. Goswami, and Y. Mitsufuji, "MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation," in *Proc. Int. Workshop on Acoustic Sign. Enhancement*, pp. 106-110, Sep. 2018.

[16] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 14, no. 4, pp. 1462-1469, 2006.

[17] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 3431-3440, 2015.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 770-778, 2016.

[19] F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, and N. Ono, "The 2016 signal separation evaluation campaign," in *Proc. Int. Conf. Latent Variable Anal. and Sign. Separation,* pp. 323-332, 2017.

[20] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. P. Bello, "MedleyDB: A multitrack dataset for annotation-intensive MIR research," in *Proc. Int. Soc. Music Info. Retrieval Conf.*, Taipei, Taiwan, Oct. 2014.

[21] Z. Rafii, A. Liutkus, F.-R. Stter, S. I. Mimilakis, and R. Bittner, "The MUSDB18 Corpus for Music Separation," Dec. 2017.

[22] H. Kim, J. Kim, and J. Park, "Music-speech separation based background music identification in TV programs," in *Proc. HCI Korea*, pp. 1158-1161, 2019.

[23] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis, "mir_eval: A transparent implementation of common MIR metrics," in *Proc. Int. Soc. Music Info. Retrieval Conf.*, 2014.

김 혜 미 (Hyemi Kim)

2004년 2월 : 부산대학교 전자전기컴퓨터공학부 학사
2006년 2월 : 한국과학기술원 전기 및 전자공학부 석사
2006년 2월~현재 : 한국전자통신연구원 선임연구원
<관심분야> 오디오 신호처리, 기계학습, 컴퓨터비전
[ORCID:0000-0002-3446-3498]

허 운 행 (Woon-Haeng Heo)

2015년 2월 : 충북대학교 전자공학부 학사
2017년 2월 : 충북대학교 제어로봇공학전공 석사
2017년 3월~현재 : 충북대학교 제어로봇공학전공 박사 과정
<관심분야> 오디오 신호처리, 기계학습, 음성인식
[ORCID:0000-0003-3303-9000]

김 정 현 (Junghyun Kim)

1999년 2월 : 전남대학교 전산학과 졸업
2001년 2월 : 전남대학교 전산학과 석사
2001년 3월~현재 : 한국전자통신연구원 책임연구원
<관심분야> 오디오 신호처리, 기계학습
[ORCID:0000-0002-7296-2968]

박 지 현 (Jihyun Park)

1999년 2월 : 서강대학교 컴퓨터학과 석사
2005년 2월 : 충남대학교 컴퓨터공학과 박사
1999년 5월~현재 : 한국전자통신연구원 책임연구원

<관심분야> 저작권보호, 신호처리, 기계학습
[ORCID: 0000-0001-6213-597X]