딥러닝 기반의 음성/오디오 기술

Speech/Audio Processing based on Deep Learning

□ 이영한 / KETI

Ⅰ. 서 론

인간의 두뇌를 모델링하는 뉴럴 네트워크 연구는 1940년대 신경 세포의 모델링부터 시작하여 현재까지 다양한 기술이 축적되어 왔다. 특히 back-propagation이 제안된 이후에 multilayer perceptron에 대한 훈련이 가능해지면서 뉴럴 네트워크는 큰 관심을 받았다. 하지만 layer를 쌓을수록 성능이 향상되기보다 local minima에 빠져 성능이 오히려 낮아지는 경우가 보고되면서 한동안 뉴럴 네트워크는 침체기를 맞이하였다. 하지만 2006년 layer를 쌓더라도 local minima에 빠지지 않고 성능이 향상될 수 있는 DBN, RBM 개념을 시작으로 다계층 구조에서도 훈련이 가능한 방법들이 소개되면서 다시 뉴럴 네트워크가 주목 받기 시작했다[1-3]. 특히 그 시작은 음성인식이었다[4]. 즉, DBN 구조의 딥러닝 기술이 음성 인식에 활용하면서 기존의 GMM-HMM framework에서 가지고 있던 성능의 한계를 넘어섰다. 특히 2012년에 ILSVRC에서 이미지 분류 기술에 CNN 기반의 deep learning이 적용되면서 과년도 성능은 물론이고 당해 2위와도 상당한 격차를 나타내면서 1위를 달성하는 Alex-net이 소개되면서 다양한 분야에서 deep learning에 대한 연구가 진행되고 있다[5]. 음성/오디오 분류/검증/인식, 이미지 검색/분류/분할, 객체 검출, 이미지 캡셔닝, 동영상 검색/분류 등 다양한 분야에서 deep learning을 적용한 예가 소개되고 있으며 대부분의 연구에서 state-of-the-art의 성능을 보이며 연구를 이끌고 있다[3-6].

본 고에서는 위에서 설명한 다양한 연구 분야 중에서 음성/오디오 분석에서의 딥러닝 적용 사례를

※ 본 논문은 미래창조과학부 SW컴퓨팅산업원천기술개발사업 (과제번호 R0190-16-1115)을 지원받아 수행한 결과입니다.

소개한다. 이에 앞서 음성/오디오 분석에 사용되는 기본 딥러닝 구조인 RNN 구조에 대해 설명한다. 이후 음성과 오디오 처리에서의 딥러닝 사례를 소개하고 각 사례가 가지는 의미를 정리한다.

II. 음성/오디오 분석에 사용되는 딥러닝 기술

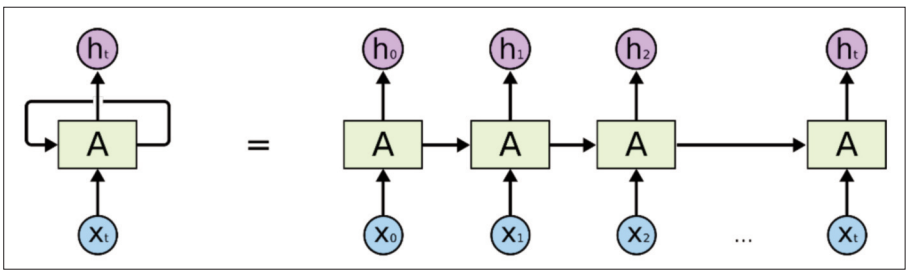
1. Recurrent Neural Network

음성/오디오 분석은 영상 처리와 다르게 원 데이터의 형식이 일반적으로 1차원 데이터라는 점과, 시계열이라는 특징을 가지고 있다. 따라서 시계열 처리를 위한 딥러닝 기법이 음성/오디오 분석에 많이 활용되고 있다. 초기에는 입력 신호에 현재 데이터뿐만 아니라, 과거 및 미래 데이터를 결합하여 overlap-shift 방식으로 처리하면서 DBN 구조로 분석에 활용하였다[7]. 하지만 최근에는 Recurrent Neural Network(RNN) 구조를 이용한 시계열 처리 방식에 딥러닝을 적용하는 형태로 연구가 많이 진행되고 있다. RNN 구조는 〈그림 1〉과 같이 기존의 hidden layer에서 loop가 추가된 형태를 의미한다. 즉, 〈그림 1〉의 오른쪽과 같이 unfolding하여 설명이 가능한데, 과거에 입력된 신호가 입력 신호의 형태로 영향을 주는 것이 아니라, hidden layer를 통해서 영향을 주는 형태이다. RNN은 이미 1980년대에 제안된 구조였지만 딥러닝 연구를 만나면서 다시 조명 받고 있다. 특히, 음성인식뿐만 아니라, 언어처리, sequence-to-sequence 등의 연구에서 채택되어 활용되고 있다.

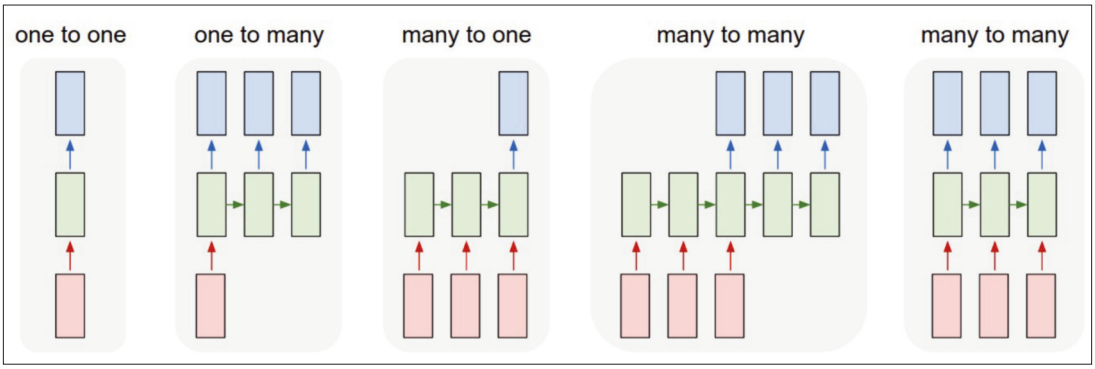
RNN이 다양한 연구에 활용될 수 있는 이유는 〈그림 2〉에서 보듯이 입력과 출력의 관계를 통해 다양한 응용분야로 활용될 수 있기 때문이다. 특히, 입력 신호를 X_0에만 입력하더라도 hidden layer를 통해서 X_1 이후의 처리에서도 출력을 추출할 수 있다. 또한 loss function의 적용 범위에 따라 출력 결과도 다양한 형태로 처리가 가능하다.

기본 뉴럴 네트워크는 〈그림 2〉(a)의 형태로 볼 수 있으며, 〈그림 2〉(b)의 경우 이미지 캡셔닝 등의 연구, 〈그림 2〉(c)의 형태는 행동 인지, 화자 인식 등의 연구, 그리고 〈그림 2〉(d),(e)는 기계 번역이나 음성 인식 등의 다양한 연구에 채택되어 활용할 수 있다는 장점을 가지고 있다.

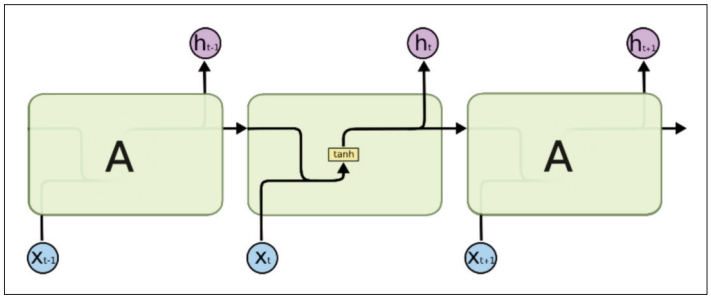
기본 RNN의 내부 구조는 〈그림 3〉와 같은 형태로 되어 있다. 즉, 상위 layer로 전달되는 출력을 다음 시간대에서 입력과 동시에 받아 처리하는 구조이다. RNN 구조는 Long-term dependencies 측



〈그림 1〉 RNN 개념도[8]



〈그림 2〉 RNN 활용예 (a) one-to-one, (b) one-to-many, (c) many-to-one, (d) many-to-many, (e) synced many-to-many[9]

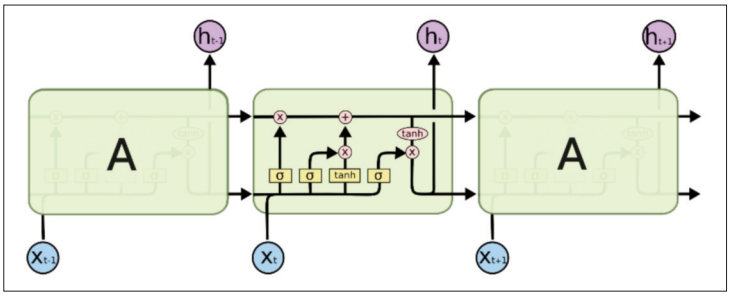


〈그림 3〉 Simple RNN 구조도[8]

면에서 이론상으로는 시간상 멀리 떨어진 내용도 잘 모델링 할 수 있어야 하지만 실제로는 오래된 과거 내용은 처리하지 못하여 성능이 떨어진다는 단점이 있다. 이는 gradient vanishing에 의해 발생하는 현상이며 구체적으로는 back-propagation through time을 연산하는데 있어 gradient가 수렴하는 것으로 인해 발생한다.

2. Long Short-term Memory(LSTM)

LSTM은 gradient vanishing을 방지하기 위해 제안된 기술 중 하나이다. 이는 1997년에 이미 제안된 구조로 현재 기본 구조 외에 다양한 변형 LSTM도 연구가 많이 되고 있다. 기본적인 LSTM의 구조는 〈그림 4〉와 같다. RNN과 비교를 하면 hidden layer 내에 메모리 기능을 넣는 동시에 메모리를 조절(쓰기/지우기/출력하기)할 수 있도록 하며 이를 훈련을 통해서 얻자는 것이 기본 아이디어이다. 실질적으로 hidden layer의 출력이 다음 시간대로 전달되는 RNN과 비교하여 별도의 정보를 추가적으로 전달하면서 처리할 수 있도록 하였다. 일반적으로 LSTM은 기존의 RNN과 비교하여 Long-term dependencies 문제에 강인한 것으로 알려져 있으며 훈련하는데 시간과 데이터가 더 많이 필요로 하지만 풍부한 데이터가 확보된 상황에서는 향상된 성능을 보이는 것으로 알려져 있다.



〈그림 4〉 LSTM 구조도[8]

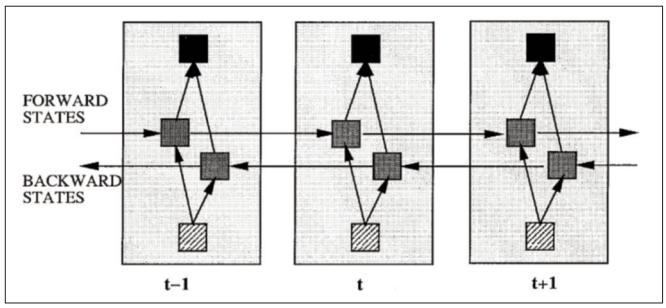
3. Bidirectional RNN

일반적인 RNN 구조에서는 과거의 정보를 현재와 미래의 처리를 위해 활용하는 구조로 되어 있다. 하지만 특정 연구 분야에서는 미래의 정보가 현재 및 과거의 정보를 처리하는데 도움을 줄 수 있기도 하다. 예를 들면, 어순이 다른 기계번역을 하는 연구를 들 수 있다. 이러한 특성을 고려하기 위해 backward directional RNN 방식이 제안되었으며, 또한 forward directional RNN 방식과 혼합하여 사용하는 bidirectional RNN 방식이 제안되었다. 〈그림 5〉는 bidirectional RNN에 대한 구조에 대해 묘사하고 있다. 특징적으로는 forward layer와 backward layer를 분리하여 다계층으로 사용하며, 최종 출력을 연산하는 layer에서만 병합하여 사용한다. 일반적으로 기존의 forward directional RNN보다 성능이 향상되는 것으로 알려져 있지만, 구조상 첫 입력에 대한 출력 역시 최종 입력을 필요로 하기 때문에 알고리즘 지연이 발생한다는 단점이 존재한다.

Ⅲ. 딥러닝 기반의 음성 분석 사례

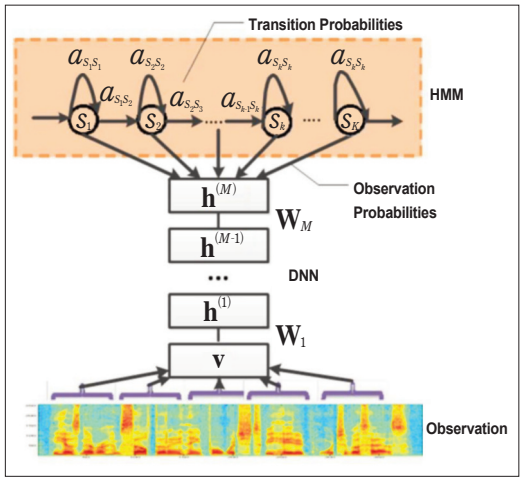
1. 음성 인식

음성 인식은 음성/오디오 분석의 대표적인 응용

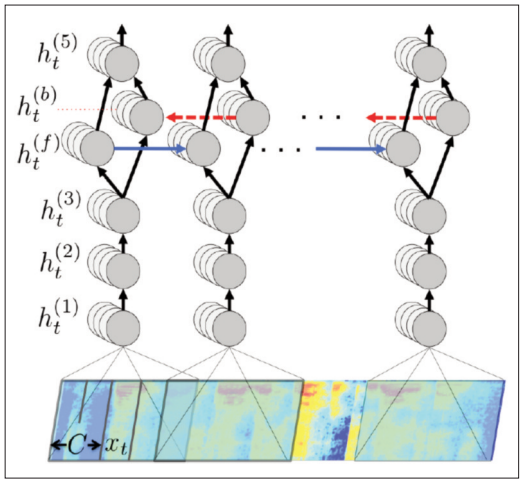


〈그림 5〉 Bidirectional RNN 구조도[10]

분야로 다양한 방식을 딥러닝 기법이 소개되기 전에는 GMM-HMM 기반의 음성 인식 기술이 주를 이루었다. 2006년 딥러닝이 소개된 이후, GMM-HMM에서 DBN-HMM 기반의 음성인식 기술이 소개되면서 음성인식 성능이 급격하게 향상되었다[7,11,12]. 특히 대용량 음성 데이터의 확보와 맞물리면서 DBN-HMM 기반의 기술을 넘어서 deep speech와 같이 spectrogram에서 캐릭터 단위로 인식하는 기술까지 소개되었다[13,14]. 특히 deep speech는 end-to-end 딥러닝이 적용되었다는 점에서 시사하는 바가 크다. Deep speech[13]는 2014년 Baidu에 의해 소개되었다. 세부적으로는 spectrogram을 입력으로 하여 CNN과 LSTM 기반의 딥러닝 모듈을 활용하여 개발되었다. 특히 bidirectional LSTM을 적용하여 성능을 향상시켰지만 backward direction 때문에 발생하는 알고리즘 지연으로 인해 실시간 서비스 제공에 단점이 존재했다. 2015년 소개된 deep speech v2[14]에서는 bidirectional LSTM 대신 row convolutional layer를 활용하여 과거 일정 시간의 정보를 활용하는 동시에 알고리즘 지연을 조절할 수 있도록 하여 실시간 서비스에 한층 다가섰다.



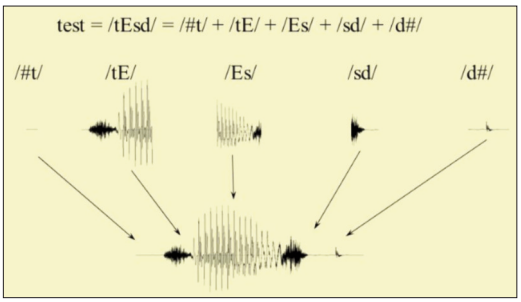
〈그림 6〉 DBN-HMM 기반 음성인식 구조도[7]



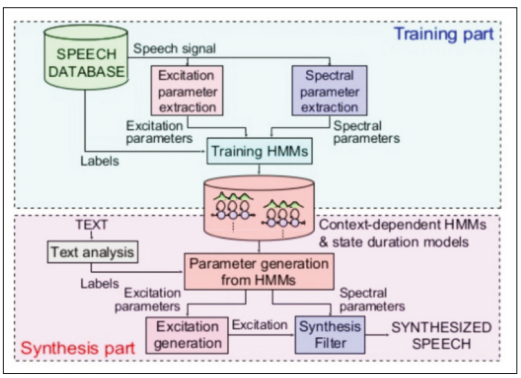
〈그림 7〉 End-to-end 딥러닝 기반 음성인식[13]

2. 음성 합성

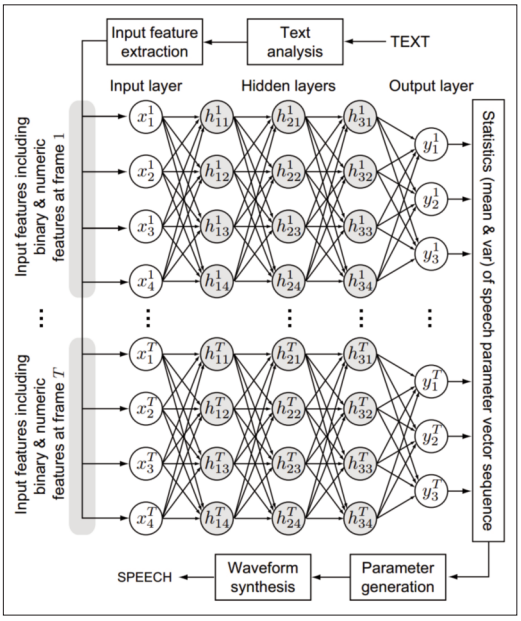
음성 합성은 음성 인식의 반대되는 개념으로 문자를 음성 신호로 변환하는 기술을 의미한다. 일반적으로 unit-selection concatenation 방식과 synthesis 방식으로 구분할 수 있다. Unit-selection 방식은 일정 단위의 음소 또는 단어를 이어 붙여서 단어나 문장을 생성하는 기술이다. 유닛 DB의 크기에 따라 음질이 좌우되며 상대적으로 synthesis 방식에 비해 음질이 좋은 것으로 알려져 있다. 반면, synthesis 방식은 음소에 해당하는 신호를 LPC 계열의 vocoder를 활용하여 합성하는 방식이다. 일반적으로 HTS(HMM-based speech synthesis system) toolkit[15]을 많이 활용하며 작은 용량으로 MOS 3.0 이상의 음질을 제공하는 것으로 알려져 있다. 음성 합성에서 딥러닝 기술



〈그림 8〉 Unit-selection 기반의 음성합성[16]



〈그림 9〉 HMM 기반의 음성합성[17]

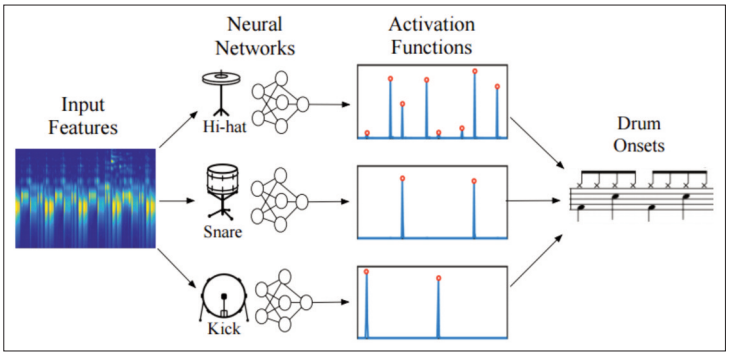


〈그림 10〉 DNN 기반의 음성합성[18]

은 synthesis 방식에 적용된 사례가 소개되었다. HTS 개발자인 Heiga zen의 ICASSP 2014, 2015 논문에 의하면 DNN 구조를 이용하여 학습을 하였고 prosody trajectory를 보면 기존의 HMM 기반의 방식보다 딥러닝을 적용했을 때 향상된 것을 확인할 수 있다. 음질 평가에 대한 결과 역시 DNN 구조에 대해 선호도가 높은 것으로 나타났다[18].

최근에는 음성을 딥러닝을 이용하여 sample 단위로 생성하는 구조가 제안되었다. Google Deepmind에서 제안한 Wavenet와 캐나다 몬트리올 대학 연구팀이 제안한 sampleRNN이 있다[19-20]. Wavenet은 RNN 구조가 아닌 Causal convolutional layer라는 개념으로 과거의 정보를 이용할 수 있는 구조를 제안하였다. HTS와 음질 평가를 진행하였는데, 음질 선호도 테스트에서 기존의 방식인 HTS에 비해 향상된 음질을 제공하는 것을 확인하였다. SampleRNN은 각 오디오 샘플 단위의 생성이 가능하다는 점이 차별점이며 tier라는 개념으로 높은 tier의 구조일수록 recurrent 성분이 아닌 입력 성분에 대해 super frame의 개념으로 접근할 수 있다는 점이 특징이다. 논문에서는 자체 구현한 Wavenet과의 음질 선호도 평가를 진행하였는데, 제안한 sampleRNN이 Wavenet보다 향상된 품질을 제공하는 것으로 나타났다. 두 방식의 단점으로는 소리를 sample 단위로 생성하기 때문에 연산량이 높다는 점이다.

Wavenet이나 sampleRNN은 음성 합성을 목적으로 개발된 딥러닝 기법은 아니지만, 소리 생성에 대한 모델링을 통해서 음성 합성의 새로운 접근 방법으로 활용할 수 있다는 점에서 큰 관심을 가지고 있다.

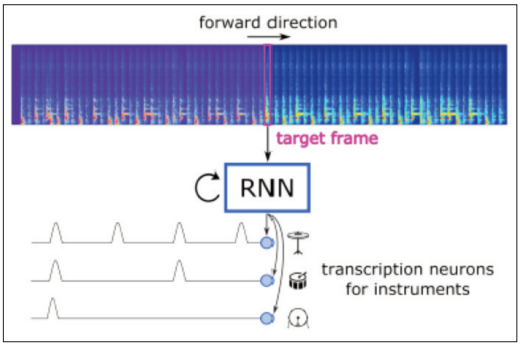


〈그림 11〉 독립모델 기반의 드럼 전사 구조도[21]

Ⅳ. 딥러닝을 활용한 오디오 분석 사례

1. 드럼 전사(Transcription) 기술

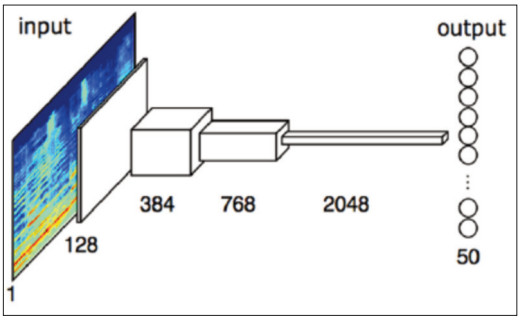
드럼 전사 기술은 드럼에 사용되는 kick, snare 그리고 hi-hat의 타격 시점을 찾는 기술이다. 이 기술은 나중에 음악의 박자를 찾는데 활용되며 장르 분류에도 큰 영향을 주는 정보이기 때문에 음악 분류 기술 중에 중요한 분야라 할 수 있다. 2016년 ISMIR 학회에서 〈그림 11〉, 〈그림 12〉와 같이 동일한 주제에 대해 논문이 투고되었다. 두 연구 모두 RNN의 구조를 활용하여 제안되었다. 특히 forward direction RNN뿐만 아니라, Backward, Bi-directional RNN 등 다양한 구조에 대해 진행한 결과가 공유되었다. 먼저 1논문에서는 입력신호를 이용하여 kick, snare, hi-hat에 대해 각각 다른 모델을 구성하여 처리하였으며 2논문에서는 단일 모델로 하여 multi-label 구조로 구성하였다. 성능은 두 연구 모두 기존의 방식보다 향상된 결과를 도출하였다. 다만 LSTM이나 GRU를 사용하지 않고 Simple RNN을 사용하여 연구를 진행하였다는 점이 특징이다.



〈그림 12〉 단일모델 기반의 드럼 전사 예시[22]

2. 자동 태깅 기술

자동 태깅 기술이란 입력된 음악에 대해 장르 정보 및 분위기 등과 같은 meta-data를 찾아주는 기술이다. 2016년 CNN을 기반으로 자동 태깅 기술을 구현하는 방법이 제안되었다[23]. 제안된 자동 태깅 기술의 알고리즘 대표도는 〈그림 13〉와 같다. 일반적으로 오디오 신호 처리에서 RNN 계열의 딥러닝을 사용하는데 비해 본 논문은 CNN을 기반으로 한다는 점이 특징이다. 다양한 길이를 가지는 오디오



〈그림 13〉 CNN 기반의 자동 태깅 기술 구조도[23]

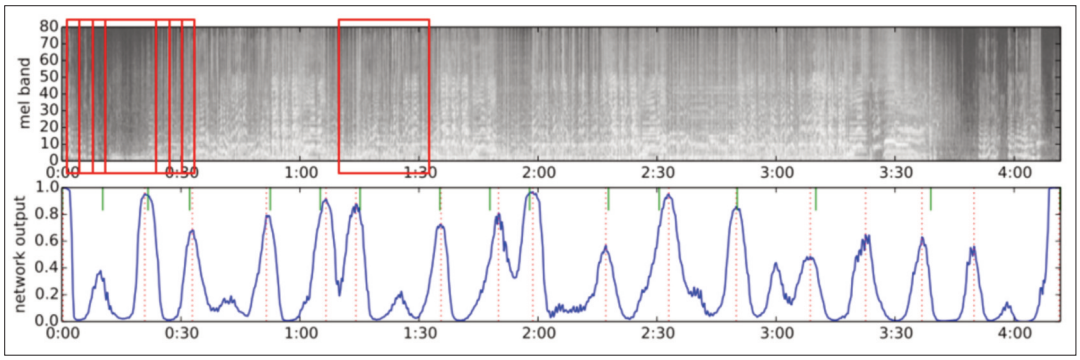
신호의 특성을 풀기 위해 제안된 연구에서는 입력 mel-spectrogram에서 가운데 특정 길이만큼의 frame을 활용하여 CNN의 입력으로 사용하였다. 또한 일반적인 분류 시스템이 단일 라벨을 기준으로 훈련을 하였다면 제안된 연구에서는 멀티 라벨을 기준으로 하여 훈련을 하였다는 것이 특징이다.

3. 오디오 분할(Segmentation) 기술

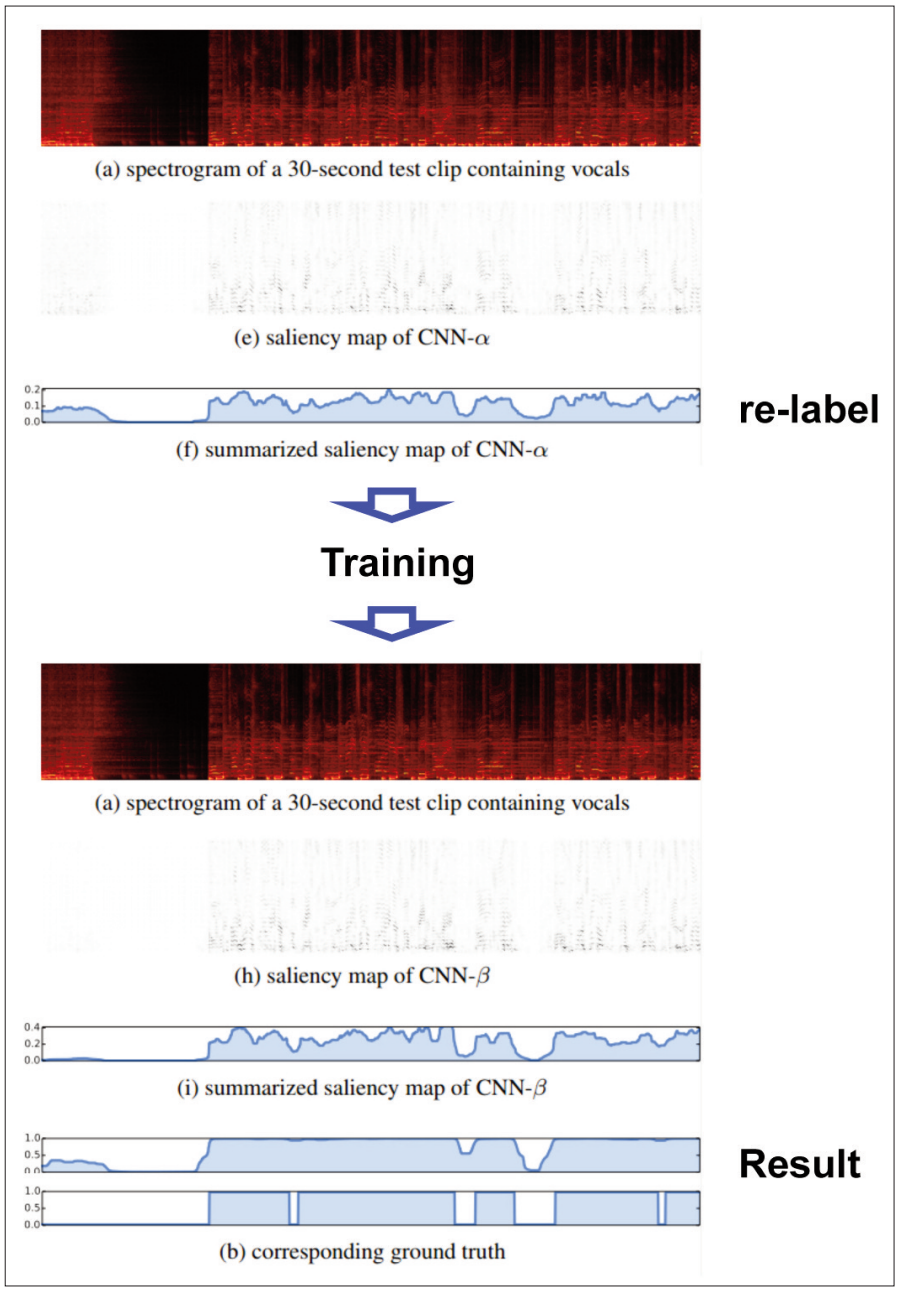
오디오 분할 기술은 주어진 오디오에 대해 시간 단위로 전주/간주 등과 같은 단위로 컨텐츠를 분할하는 기술이다. 구체적으로는 경계 검출을 위한 boundary detection과 분할된 단위의 라벨을 분류하는 label classification으로 구성된다. 2014년 boundary detection을 위한 방법으로 CNN을 이용한 기법이 소개되었다[24]. 제안된 방식에서는 자동 태깅 기법과 마찬가지로 mel-spectrogram을 입력으로 사용하였고, frame별 경계 유무를 판별하기 위해서 sliding-window 기법으로 결과를 도출하였다. 입력된 mel-spectrogram 대비 boundary detection 결과는 〈그림 14〉와 같다.

4. 보이스 핀포인트(pinpoint)

보이스 핀포인트 기법은 1차적으로 weak label되어있는 database를 활용하여 hard label인 음성 구간을 검출하고 spectrogram상에서 음성 구역을 검출하는 기술을 말한다. 여기서 weak label이란 음성 구간에 대해 label되어 있는 데이터셋이 아닌 파일 단위에서 음성의 유무로만 label이 되어 있는 데이터셋을 의미한다. 즉, 직접적인 목표인 음성 구간에 대해 label되어 있진 않지만 그보다 rough하게 label되어 있는 경우를 의미한다. 2016년 CNN 기법을 활용하여 검출하는 기술이 소개되었다[25]. 먼저 CNN을 weak



〈그림 14〉 CNN 기반의 boundary detection 결과 예시[24]



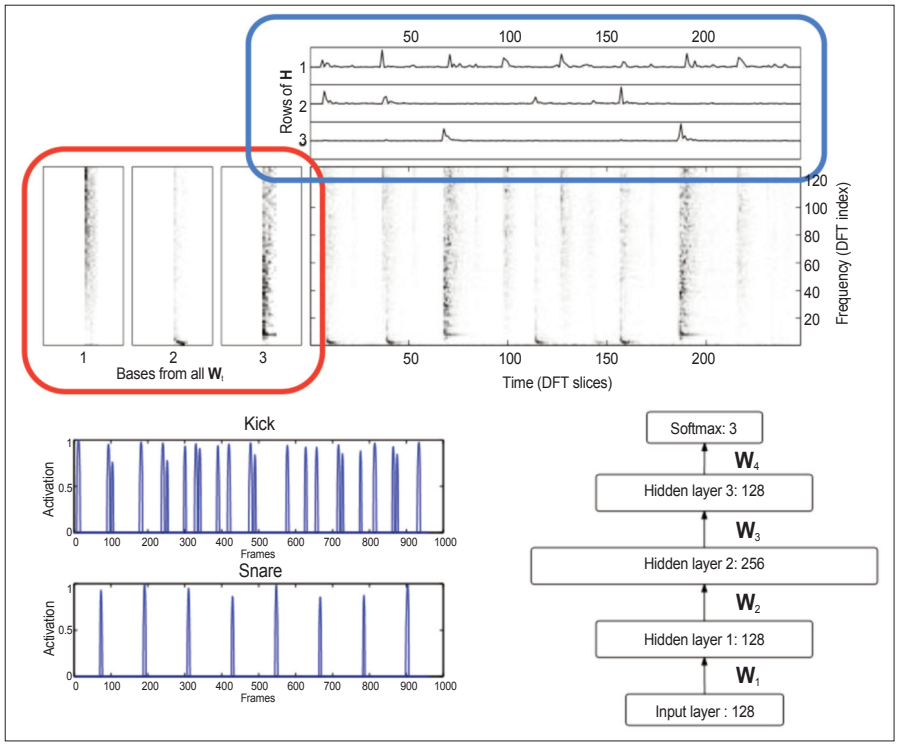
〈그림 15〉 보이스 핀포인트 처리 순서도[25]

label을 이용하여 sliding window 방식으로 훈련을 한다. 이렇게 할 경우 〈그림 15〉(f)에서 보는 것과 같이 정확하진 않지만 음성이 없는 구간에서 confidence 수치가 낮아지는 것을 확인할 수 있다. 이

는 음성이 없는 파일에서 훈련된 영향이다. 이렇게 나온 결과치를 label로 활용하는 과정을 반복함으로써 제안된 연구에서는 음성 구간을 검출하였다. 이와 함께 guided back-propagation을 이용하여 〈그림 15〉와 같이 spectrogram 상에서 음성 구역만 찾는 것으로 활용하였다. Weak label을 통해 hard label을 추정할 수 있다는 점이 특징인데, 실제 label을 만드는 작업에 많은 시간과 비용이 소비되는 점은 감안하면 제안된 연구가 가지는 시사점은 높다고 할 수 있다.

5. 오디오 분리(Separation) 기술

오디오 분리 기술은 시간축에서 정보를 나누는 오디오 분할과 달리 이미 시간축에서 혼합되어 있는 신호를 특성별로 분리하는 기술을 의미한다. 예를 들면 음성 제거, 악기 추출 등이 오디오 분리 기술에 속한다고 볼 수 있다. 2015년에는 NMF 기반의 오디오 분리 기술이 소개되었다[26]. 하지만 NMF의 경우 오디오 신호 성분의 특성별 분리는 가능하지만 분리된 성분이 악기나 음성에 해당하는지 분류하지 못한다는 단점이 있다. 이러한 점을 해결하기 위해 2014년 관련 연구가 제안되었다. 전체 구조는 〈그림 16〉과 같고 각 latent source별 분류를 통해서 최종적으로 snare 와 kick의 오디오 신호를 분리하는 연구다. NMF의 base를 이용하여 분류하는 방식을 채택하였으며 딥러닝 구조로는 DBN 구조로 하여 구현하였다.



〈그림 16〉 딥러닝을 활용한 NMF 기반의 음성 분리 성분 분류[26]

특히 제안된 기법의 경우 기존에 사용하던 방식과 비교하여 특징 벡터 추출 등의 과정이 딥러닝 구조 내부로 흡수되면서 보다 간단하게 구현되었다는 점이 특징이다.

V. 결 론

본 고에서는 딥러닝 기반의 음성/오디오 분석 기술에 대해 살펴 보았다. 기본적으로 음성/오디오 분석에 사용되는 딥러닝 구조에 대해 살펴보았으며 이를 활용한 다양한 분야의 예시를 살펴보았다.

이미지/영상에 비해 음성/오디오와 관련된 딥러닝 연구는 상대적으로 접근하기 쉽지 않다. 먼저 음성의 경우 각 언어별 기본적으로 사용하는 언어처리가 필요하기 때문에 변환이 필요하다. 또한 한국어 데이터베이스의 경우 공개된 데이터 양이 부족하거나 대부분 비공개 데이터베이스이기 때문에 쉽게 접근하기 어렵다. 오디오의 경우에는 저작권 문제로 인해 데이터베이스의 공유가 어렵다는 한계를 지니고 있다. 따라서 label 파일을 공유하더라도 데이터베이스는 자체적으로 구축해야 하는 단점이 있다. 이를 해결하기 위해 internet archive을 사용하거나 raw-audio가 아닌 feature level에서의 데이터를 생성하여 공유하는 방안으로 접근 중이지만, 여전히 상용 오디오와 특성이 다르거나 feature에 의한 손실이 존재하기 때문에 근본적인 해결책이 되지 못하고 있다.

그럼에도 음성은 기본적인 커뮤니케이션 수단이라는 점과 컨텐츠 소비 시장에서 오디오의 비중이 여전히 높기 때문에 음성/오디오 분석에 대한 연구 및 요구는 지속될 것으로 예상된다. 특히 기존의 기법보다 향상된 성능을 확보하기 위해 딥러닝을 적용하는 연구는 많은 관심을 받을 것으로 예상된다.

참 고 문 헌

[1] Hinton, Geoffrey E. et al. “Reducing the dimensionality of data with neural networks.” Science, (2006)

[2] Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. “A fast learning algorithm for deep belief nets.” Neural computation 18.7 (2006): 1527-1554.

[3] Bengio, Yoshua, et al. “Greedy layer-wise training of deep networks.” Advances in neural information processing systems 19 (2007): 153.

[4] Hinton, Geoffrey, et al. “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups.” IEEE Signal Processing Magazine 29.6 (2012): 82-97.

[5] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. “Imagenet classification with deep convolutional neural networks.” Advances in neural information processing systems. (2012).

[6] Donahue, Jeffrey, et al. “Long-term recurrent convolutional networks for visual recognition and description.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015).

[7] Dahl, George E., et al. “Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition.” IEEE Transactions on Audio, Speech, and Language Processing 20.1 (2012): 30-42.

[8] Understanding LSTM Networks, http://colah.github.io/posts/2015-08-Understanding-LSTMs/

[9] The Unreasonable Effectiveness of Recurrent Neural Networks, http://karpathy.github.io/2015/05/21/rnn-effectiveness/

[10] Schuster, Mike, and Kuldip K. Paliwal. “Bidirectional recurrent neural networks.” IEEE Transactions on Signal Processing, IEEE, (1997)

[11] Mikolov, Tomás, et al. “Strategies for training large scale neural network language models.” Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on. IEEE, (2011).
[12] Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton. “Speech recognition with deep recurrent neural networks.” 2013 IEEE international conference on acoustics, speech and signal processing. IEEE, (2013).
[13] Hannun, Awni, et al. “Deep speech: Scaling up end-to-end speech recognition.” arXiv preprint arXiv:1412.5567(2014).
[14] Amodei, Dario, et al. “Deep speech 2: End-to-end speech recognition in English and mandarin.” arXiv preprint arXiv:1512.02595(2015).
[15] A.W. Black, H. Zen, K. Tokuda, ”Statistical parametric speech synthesis.” 2007 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, (2007).
[16] Speech Synthesis, http://slideplayer.com/slide/3148265/
[17] http://www.slideshare.net/danilosoba1/statistical-parametric-speech-synthesis-heiga-zen
[18] Zen, Heiga, Andrew Senior, and Mike Schuster. “Statistical parametric speech synthesis using deep neural networks.” 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, (2013).
[19] van den Oord, Aäron, et al. “Wavenet: A generative model for raw audio.” arXiv preprint arXiv:1609.03499(2016)
[20] Soroush Mehri, et al. “SampleRNN: An Unconditional End-to-End Neural Audio Generation Model.” https://openreview.net/forum?id=SkxKPDv5, under review on ICLR 2017.
[21] Southall, Carl, Ryan Stables, and Jason Hockman. “AUTOMATIC DRUM TRANSCRIPTION USING BI-DIRECTIONAL RECURRENT NEURAL NETWORKS.” Proceedings of the International Society for Music Information Retrieval Conference (ISMIR). (2016).
[22] Vogl, Richard, Matthias Dorfer, and Peter Knees. “RECURRENT NEURAL NETWORKS FOR DRUM TRANSCRIPTION.” Proceedings of the International Society for Music Information Retrieval Conference (ISMIR). (2016).
[23] Choi, Keunwoo, George Fazekas, and Mark Sandler. “Automatic tagging using deep convolutional neural networks.” Proceedings of the International Society for Music Information Retrieval Conference (ISMIR). (2016).
[24] Schlüter, Jan, Karen Ullrich, and Thomas Grill. “Structural segmentation with convolutional neural networks mirex submission.” 10th running of the Music Information Retrieval Evaluation eXchange (MIREX 2014) (2014).
[25] Schlüter, Jan. “Learning to pinpoint singing voice from weakly labeled examples.” Proceedings of the International Society for Music Information Retrieval Conference (ISMIR). (2016).
[26] Leimeister, Matthias. “Feature learning for classifying drum components from nonnegative matrix factorization.” Audio Engineering Society Convention 138. Audio Engineering Society, (2015).

필 자 소 개

이 영 한

– 2005년 2월 : 광운대학교 전자공학과 학사
– 2007년 2월 : 광주과학기술원 정보통신공학과 석사
– 2011년 8월 : 광주과학기술원 정보통신공학부 박사
– 2011년 7월 ~ 2014년 12월 : LG전자기술원 선임연구원
– 2015년 1월 ~ 현재 : 전자부품연구원 근무
– 주관심분야 : 음성/오디오 신호처리, 머신러닝