일반논문 (Regular Paper)
방송공학회논문지 제23권 제6호, 2018년 11월 (JBE Vol. 23, No. 6, November 2018)
https://doi.org/10.5909/JBE.2018.23.6.855
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

실생활 음향 데이터 기반 이중 CNN 구조를 특징으로 하는 음향 이벤트 인식 알고리즘

서 상 원[a)], 임 우 택[a)], 정 영 호[a)‡], 이 태 진[a)], 김 휘 용[a)]

Dual CNN Structured Sound Event Detection Algorithm Based on Real Life Acoustic Dataset

Sangwon Suh[a)], Wootaek Lim[a)], Youngho Jeong[a)‡], Taejin Lee[a)], and Hui Yong Kim[a)]

요 약

음향 이벤트 인식은 다수의 음향 이벤트가 발생하는 환경에서 이를 인식하고 각각의 발생과 소멸 시점을 판단하는 기술로써 인간의 청각적 인지 특성을 모델화하는 연구다. 음향 장면 및 이벤트 인식 연구 그룹인 DCASE는 연구자들의 참여 유도와 더불어 음향 인식 연구의 활성화를 위해 챌린지를 진행하고 있다. 그러나 DCASE 챌린지에서 제공하는 데이터 세트는 이미지 인식 분야의 대표적인 데이터 세트인 이미지넷에 비해 상대적으로 작은 규모이며, 이 외에 공개된 음향 데이터 세트는 많지 않아 알고리즘 개발에 어려움이 있다. 본 연구에서는 음향 이벤트 인식 기술 개발을 위해 실내외에서 발생할 수 있는 이벤트를 정의하고 수집을 진행하였으며, 보다 큰 규모의 데이터 세트를 확보하였다. 또한, 인식 성능 개선을 위해 음향 이벤트 존재 여부를 판단하는 보조 신경망을 추가한 이중 CNN 구조의 알고리즘을 개발하였고, 2016년과 2017년의 DCASE 챌린지 기준 시스템과 성능 비교 실험을 진행하였다.

Abstract

Sound event detection is one of the research areas to model human auditory cognitive characteristics by recognizing events in an environment with multiple acoustic events and determining the onset and offset time for each event. DCASE, a research group on acoustic scene classification and sound event detection, is proceeding challenges to encourage participation of researchers and to activate sound event detection research. However, the size of the dataset provided by the DCASE Challenge is relatively small compared to ImageNet, which is a representative dataset for visual object recognition, and there are not many open sources for the acoustic dataset. In this study, the sound events that can occur in indoor and outdoor are collected on a larger scale and annotated for dataset construction. Furthermore, to improve the performance of the sound event detection task, we developed a dual CNN structured sound event detection system by adding a supplementary neural network to a convolutional neural network to determine the presence of sound events. Finally, we conducted a comparative experiment with both baseline systems of the DCASE 2016 and 2017.

Keyword : Machine learning, Deep learning, Audio signal processing, Sound event detection, Dataset

Ⅰ. 서 론

인간은 청각 정보를 통해 주변에서 일어나는 사건들을 인지하며, 이를 다른 감각 정보와 연계해 판단을 도출한다. 기계학습 분야에는 이러한 인간의 청각적 인지 특성을 모델화하기 위한 다양한 연구가 진행되고 있으며, 그중 음향 이벤트 인식(sound event detection) 기술은 최근 빠르게 성장하고 있는 분야이다[1-4]. 특히 음향 이벤트 인식 기술은 청력이 저하된 사람들을 위한 공간 지각 서비스, 자율 주행 차량 개발 그리고 스마트 시티 구축을 비롯한 많은 응용 분야에 큰 영향을 미칠 수 있으므로 계속하여 주목을 받고 있다.

음향 장면 및 이벤트 인식과 관련된 다양한 논의를 진행하는 연구 그룹인 DCASE(Detection and Classification of Acoustic Scenes and Events)에 따르면, 음향 이벤트 인식이란 다수의 음향 이벤트가 발생하는 환경에서 알고리즘이 이벤트를 인식하고 각각의 발생(onset)과 소멸(offset) 시점을 판단할 수 있는지를 평가하는 과제이다. 이 과제는 2013, 2016 그리고 2017년에 진행된 DCASE 챌린지[5-7]에 포함되었으며, 다양한 알고리즘들이 제안되었다. DCASE 챌린지 초기에는 비음수 행렬 분해(non-negative matrix factorization)[2] 및 가우시안 혼합 모델(gaussian mixture model)[8] 등 전통적인 기계 학습 기법을 기반으로 한 알고리즘들이 제출되었으나, 최근에는 RNN(recurrent neural network)[9]과 CNN(convolutional neural network)[10] 등 딥 러닝(deep learning) 기반의 알고리즘들이 주류를 이루고 있다. 특히 2017년에는 CNN과 RNN을 연결한 구조인 CRNN(convolutional recurrent neural network) 기반의 알고리즘[11]이 제안되었으며, 제출된 알고리즘 중 가장 우수한 음향 이벤트 인식 성능을 보였다. 하지만, 이벤트 존재에 대한 오판단이 높아 실제 활용되기에는 낮은 인식 성능을 보여 개선이 필요한 실정이다.

한편, 이미지 인식 분야에서는 인간의 분류 오차인 5~10%보다 우수한 3.6%의 오류율을 기록한 알고리즘[12]이 2015년 ILSVRC(ImageNet Large Scale Visual Recognition Challenge)[13]에 보고되었다. 이미지 인식 알고리즘이 이처럼 기계학습 분야에서 선구적인 발전을 이루어 낸 기반 중 하나에는 이미지넷(ImageNet)[14]이라는 강력한 데이터 세트의 제공이 있었다. ILSVRC에서는 이러한 이미지넷 데이터 세트 중 일부를 활용하여 1,000개의 카테고리를 갖는 120만 장의 이미지를 훈련 세트로 제공하였다. 반면 DCASE 2017 챌린지의 음향 이벤트 인식 작업을 위한 데이터 세트 규모는 여섯 종류의 이벤트 클래스에 대해 3~5분 길이의 오디오 클립을 24개 제공하며, 총 659개 인스턴스를 포함하도록 구성되어 있다. 둘 다 클래스당 인스턴스 수가 일정하지 않은 불균형 데이터세트(imbalanced dataset)이므로 클래스별 정도의 차이는 있으나, 이미지넷 챌린지는 DCASE 챌린지와 비교하여 더 큰 규모의 데이터 세트를 제공하고 있다. 따라서 추가적인 음향 데이터 세트의 개발이 필요한 상황이지만, 획득 과정상의 어려움 및 저작권 이슈 등으로 인해 공개된 데이터의 수가 매우 적다.

본 연구에서는 음향 이벤트 존재 구간에 대해 오판이 높은 state-of-the-art 알고리즘의 보완을 위해 유효한 이벤트가 존재하는 구간을 판단할 수 있는 보조 네트워크를 포함한 이중 CNN 구조를 제안한다. 또한 데이터 세트 부족 문제를 해결하고자 음향 데이터 세트 제작을 수행하여, 제안하는 음향 이벤트 인식 알고리즘의 학습과 평가에 활용하였다.

본 논문의 구성은 다음과 같다. 먼저 II 장에서는 음향 이벤트 인식 기술 개발을 위한 데이터 세트 설계, 수집 및 메타데이터 상세에 관해 설명한다. 다음으로 III 장에서는 이벤트의 존재 여부를 판단하는 보조 신경망을 추가한 이중 CNN 기반 음향 이벤트 인식 알고리즘을 제안한다. IV 장에서는 제작된 데이터 세트를 활용하여 음향 이벤트 인식 알고리즘을 평가하고, DCASE 챌린지의 2016년 및 2017년 기준 시스템과 성능을 비교한다. 마지막으로 V 장에서는 본 논문에서 진행한 연구 내용에 대한 요약 및 음향 이벤트 인식 연구의 향후 연구 방향에 관해 기술한다.

a) 한국전자통신연구원 실감AV연구그룹 (Realistic AV Research Group, Electronics and Telecommunications Research Institute)
‡ Corresponding Author : 정영호(Youngho Jeong)
E-mail: yhcheong@etri.re.kr
Tel: +82-42-860-6472
ORCID: https://orcid.org/0000-0001-9552-8593
※ This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.2017-0-00050, Development of Human Enhancement Technology for auditory and muscle support)
※ 이 논문은 2018년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No.2017-0-00050, 신체기능의 이상이나 저하를 극복하기 위한 휴먼 청각 및 근력 증강 원천 기술 개발)
· Manuscript received August 20, 2018; Revised October 24, 2018; Accepted October 24, 2018.

Ⅱ. 음향 데이터 세트

1. 음향 데이터 세트 설계

실생활 환경에서 녹음된 음향 데이터 세트는 획득 과정상의 어려움 및 저작권 이슈 등으로 인해 공개된 데이터 세트의 수가 매우 적다. 따라서 본 연구를 위해 자체적으로 실생활 환경의 음향 데이터를 획득하고 음향 이벤트에 대한 발생(onset)과 소멸(offset) 시점 및 이벤트 클래스에 대한 어노테이션(annotation)을 진행하여 음향 데이터 세트를 구축하였다[15]. 음향 데이터 관련 요구사항은 DCASE 챌린지에서 제공하는 TUT 데이터베이스[8] 규격을 참고하였으며, 본 연구에서는 수집하고자 하는 음향 이벤트 클래스를 먼저 정의한 뒤에 음향 데이터 수집을 진행하였다. 이는 음향 이벤트당 수집 개수를 평준화하여 학습 과정에서 특정 이벤트 클래스의 인식 성능이 떨어지는 것을 막기 위함이다. 수집 대상 음향 이벤트는 아래 표 1에서 보는 바와 같이 수집 환경에 따라 실내와 실외로 구분되며, 각각 16개의 클래스로 정의된다. 향후 응용을 고려하여 실내외에서 발생할 수 있는 위험 상황에 대한 이벤트 클래스를 추가하여 데이터 세트를 구성했으며, 비위험 상황 이벤트 클래스의 경우에는 DCASE 2016 및 2017 챌린지에서 제시한 이벤트 클래스를 참조하였다[5,6].

표 1. 음향 이벤트 클래스
Table 1. Sound event classes

	Event class label	Number of instances	Hazard event
Indoor	Kettle whistle	227	O
	Children crying	255	O
	Children playing	602	-
	Children shouting	94	-
	Dish cleanup	249	-
	Dish rinse sound	121	-
	Dishwasher	143	-
	Doorbell	201	-
	Drawer sound	155	-
	Drop impact sound	287	O
	Fire alarm sound	218	O
	Footfall	139	-
	Keyboard sound	95	-
	Scream	236	O
	Speech	411	-
	Water flowing	211	-
Outdoor	Bicycle idle horn	127	O
	Bird singing	504	-
	Car crash	82	O
	Car idle horn	90	O
	Car passing	173	O
	Car passing horn	130	O
	Drop impact sound	231	O
	Footfall	485	-
	Motorcycle idle horn	141	O
	Motorcycle passing horn	212	O
	Scream	122	O
	Speech	3	-
	Truck idle horn	108	O
	Truck passing	123	O
	Truck passing horn	206	O
	Wind sound	593	-

음향 신호 수집 규격은 표 2와 같이 구성되며, 본 연구에서는 향후 응용을 고려하여 바이노럴 신호뿐만 아니라 스테레오 신호도 함께 수집하였다. 수집된 음향 데이터는 3~5분 사이의 오디오 파일로 편집되어 저장되었다.

표 2. 녹음 신호 규격
Table 2. Recording signal specifications

Signal type	Binaural / Stereo
Sampling rate	44.1 kHz
Bit resolution	24 bits
Audio format	PCM WAV

2. 음향 데이터 수집 및 어노테이션 작업

실생활 음향 데이터 수집을 위해 바이노럴 및 스테레오 마이크를 사용하는 음향 데이터 수집 시스템을 구축하였다. 바이노럴 음향 신호는 인이어 마이크로폰을 사람이 직접 착용한 상태에서 수집하는 것을 기본으로 하되, 위험 요소 등으로 인해 사람이 수행하기 어려운 상황으로 판단될 경우 더미 헤드를 활용하였다. 인이어 마이크로폰으로는 Soundman OKM II Klassik/studio A3 electret in-ear microphones을 이용하며, 스테레오 음향 신호 수집에는 고감도 스테레오 마이크로폰인 Rode NT4 X/Y Stereo Condenser Microphone을 사용하였다. 각각의 마이크로폰을 통해 입력된 음향 신호는 TASCAM DR-100mkII PCM 포터블 레코더의 외부 마이크 입력을 통해 녹음되었다. 음향 데이터 수집 중에 인이어 마이크로폰을 착용한 수집자는 가급적 대화와 신체 움직임 또한 최소한으로 할 수 있도록 제한했으

며, 이는 잡음의 삽입 억제 및 향후 음향 공간정보 활용 가능성을 높이기 위한 목적이다. 또한, 음향 환경에 따른 음향 이벤트의 변동성 확보를 위해 하나에 이벤트 클래스에 대해 최소 10곳 이상의 서로 다른 장소와 시간대를 선정하여 장시간에 걸쳐 수집하였다. 가정이나 사무실 등과 같은 사생활과 관련된 음향 데이터가 수집될 수 있는 환경에서는 참여한 모든 이들로부터 개인 정보 제공 및 활용 동의서를 받고 수집을 진행하였으며, 공원, 인도 등과 같은 공공장소의 경우에는 사생활과 관련된 음향 데이터가 거의 수집되지 않기에 별도 동의서를 받지 않고 음향 데이터를 수집하였다.

어노테이션 작업은 훈련된 2명 이상의 전문가가 참여하여 진행하였으며, 생성된 메타데이터에 대한 상호 검증을 통해 작업 결과에 대한 신뢰성을 확보하였다. 또한, 어노테이션 단계에서 사생활 정보가 포함되었는지를 확인하여, 침해 가능성이 있다고 판단되는 음향 구간은 제거하였다. 구축된 음향 데이터 세트는 binaural 및 stereo에 대해 각각 총 13시간 9분 분량의 230개 파일로 구성되며, 표 1에서 정의된 음향 이벤트 클래스를 기준으로 각 클래스당 평균 254개의 음향 이벤트를 포함한다. 이는 DCASE 2017 챌린지의 음향 이벤트 인식 과제 데이터 세트가 1시간 31분 분량의 파일을 제공한 것에 비해 약 8.7배 더 큰 규모이며, 각 클래스당 평균 1.7배 정도 더 많은 인스턴스 수를 갖고 있다. 어노테이션 작업을 통해 생성되는 음향 이벤트 메타데이터의 구성 예시는 표 3과 같다.

표 3. 음향 이벤트 메타데이터 구성 예
Table 3. Examples of sound event metadata

onset	offset	event class label
2.087000	10.354000	footfall
16.977000	21.690000	footfall
26.481000	31.470000	water flowing
32.999000	40.825000	dishwashing
38.326000	38.984000	drop impact sound

III. 음향 이벤트 인식 알고리즘

1. 이중 CNN 기반 음향 이벤트 인식 알고리즘

본 장에서는 제안하는 이중 CNN 기반 음향 이벤트 인식 알고리즘을 상세히 설명한다. 본 논문에서는 일반적인 컨벌루션 신경망 구조를 주 인식 네트워크로 하면서, 이벤트 존재 여부에 대한 오검출을 줄일 수 있는 보조 네트워크를 추가하여 시스템을 구성함으로써 음향 이벤트 인식을 수행한다. 전술한 바와 같이 음향 이벤트 인식이 낮은 성능을 보이는 주된 이유 중 하나는 이벤트의 존재 여부에 대한 오판율이 높기 때문인데, 이를 확인하기 위해 본 실험에 앞서 먼저 음향 이벤트 인식 결과를 분석하였다. 분류 문제의 성능을 분석할 때는 일반적으로 오차 행렬(confusion matrix)이 주로 사용되지만, 다중 음향 이벤트 인식의 경우 2개 이상의 이벤트가 동시에 존재할 수 있어서 오차 행렬을 통한 정확한 분석이 어렵다. 따라서 본 연구에서는 다중 음향 이벤트가 존재하는 경우 중 오차 행렬을 구성할 수 없는 경우를 일부 예외 처리하여 인식 결과의 경향성을 분석하였다. 기존에 DCASE에서 제공하는 기준 시스템들을 활용해 인식 결과를 통해 분석한 결과, 실제 이벤트가 존재하지 않지만 인식 결과는 유효한 이벤트가 존재한다고 오판하거나, 실제 이벤트가 존재하나 인식 결과는 유효한 이벤트가 존재하지 않는다고 판단한 경우가, 이벤트를 인식하였으나 다른 이벤트로 잘못 인식한 경우에 비해 평균적으로 10배 이상의 오검출 빈도를 보였다. 따라서 음향 이벤트 인식 성능을 높이기 위해서는 이벤트 존재 여부에 대한 보다 명확한 판단이 전제되어야 함을 가정하고 음향 이벤트 인식 알고리즘을 제안하였다. 그림 1의 (a)는 제안하는 음향 이벤트 인식 알고리즘의 신경망 구조 및 파라미터에 대한 블록선도이며, (b)는 알고리즘이 입력된 음향 데이터에 대해 음향 이벤트 인식을 수행하는 구조에 대한 개략도이다.

본 논문에서는 44.1kHz의 입력 오디오 신호를 40ms 해석 윈도우(analysis window) 단위로 50%씩 중첩하며 각 프레임당 40개의 로그 멜 밴드 에너지를 추출하였다. 추출된 오디오 특징은 음향 이벤트에 대한 문맥정보(context information)를 반영하기 위해 현재 프레임 전후로 각 25개 프레임의 특징을 결합하여 2차원 입력 이미지로 구성하였다. 이렇게 구성된 입력 이미지는 음향 이벤트 인식을 위한 주 신경망과 음향 이벤트의 존재 여부에 관한 결과를 보완하는 보조 신경망으로 입력된다. 각 신경망의 학습 및 평가는 별도로 이루어지며 각 모델의 결과가 결합 되어 최종 판단이 수행된다.

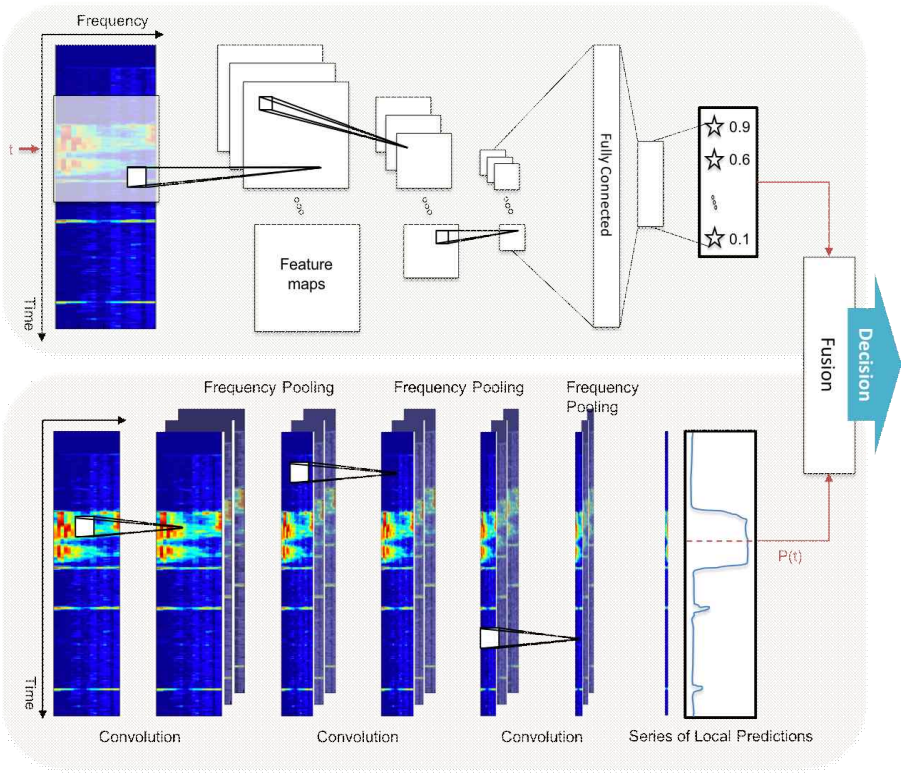

그림 1. 이중 CNN 기반 음향 이벤트 인식 알고리즘의 개략도
Fig. 1. Schematic diagram of dual CNN based sound event detection algorithm

음향 이벤트 인식을 위한 주 신경망은 총 3개의 컨벌루션 층(Convolution layer)과 2개의 전 결합 층(fully-connected layer)으로 구성된다. 각각의 컨벌루션 층은 3x3 크기의 컨벌루션 필터로 이루어진 64개의 노드로 이루어지며, 활성화 함수로는 ReLU를 사용하였다. 모든 컨벌루션 층에는 과적합을 줄이기 위한 20%의 dropout과 특징맵의 크기를 축소하는 2x2 맥스풀링(max-pooling)이 함께 수행되었다. 2개의 전 결합 층은 각각 128개의 노드와 인식 대상 클래스의 수와 동일한 노드 숫자를 가지며, 활성화 함수로는 ReLU와 sigmoid가 사용되었다. 미니 배치(minibatch)의 크기는 64, 학습률(learning rate)은 0.001, 최적화 함수로는 Adam 알고리즘[16]을 활용하여 학습되었으며, 최대 100 epochs 동안 훈련하였다. 또한, 네트워크의 과적합(over-adaptation)을 막기 위하여 10 epoch 동안 손실(loss) 값의 개선이 이루어지지 않으면 학습을 멈추도록 하였다. 이때, 검증(validation) 데이터에 대한 목푯값과 네트워크가 예측한 값 사이의 손실 값을 평가 기준(evaluation criterion)으로 사용하였다.

음향 이벤트의 존재 여부를 판단하는 보조 신경망은 총 3개의 컨벌루션 층과 하나의 시간 축 전 결합 층으로 구성된다. 컨벌루션 층은 3x3 크기의 컨벌루션 필터로 이루어진 32개의 노드로 이루어지며, 활성화 함수로는 ReLU를 사용하였다. 모든 컨벌루션 층에는 과적합을 줄이기 위한 20%의 dropout과 특징맵의 크기를 축소하는 맥스풀링이 함께 수행되었다. 이때 맥스풀링은 각 프레임별 이벤트 존재 여부에 대한 결과를 얻기 위해 시간 축의 정보는 보존한 채 주파수 축 만으로 풀링을 수행하였고, 각 시간 프레임별 1(이벤트가 존재함) 또는 0(이벤트가 존재하지 않음)을 목표로 하여 학습되었다. 보조 신경망의 결과는 특정 임계값 이상일 경우 이벤트가 존재하는 프레임이라고 결정할 수 있으며, 본 실험에서는 음향 이벤트의 존재 여부를 판단하는 보조 네트워크로 활용되었으므로 임계값을 0.2로 하였다.

따라서 최종 인식 결과를 결정할 때 임계값 이상일 경우 주 신경망의 인식 결과를 유효하다고 판단하고, 임계값 이하일 경우 주 신경망의 인식 결과에 관계없이 이벤트가 존재하지 않는다고 판단하여 실험을 수행하였다. 각 신경망의 학습 및 평가는 TensorFlow[17]를 백엔드로 사용하는 Keras[18] 프레임워크를 활용하여 구현되었다.

2. DCASE 음향 이벤트 인식 시스템

제안하는 음향 이벤트 인식 알고리즘의 우수성을 검증하기 위해 본 논문에서는 2가지의 기준 음향 이벤트 인식 알고리즘과 비교하였다. 첫 번째는 가우시안 혼합 모델(Gaussian Mixture Model, GMM)이며 두 번째는 다층 퍼셉트론(Multi-Layer Perceptron, MLP) 모델이다.

먼저 가우시안 혼합 모델은 각 클래스의 확률밀도 함수를 몇 개의 가우시안 확률 밀도 함수의 선형 결합으로 가정하는 모델이다. 가우시안 혼합 모델에서 학습할 수 있는 모수는 n번째 가우시안에 대한 가중치 및 각 가우시안의 평균과 분산이며 이러한 모수들은 EM(Expectation Maximization) 알고리즘을 활용하여 실제 데이터의 확률 분포 모델과 유사한 모델을 찾도록 학습할 수 있다. 본 실험에 활용된 가우시안 혼합 모델은 DCASE 2016의 기준 시스템[8]으로 제안된 모델을 기반으로 하였으며, 입력으로 활용되는 오디오 특징은 MFCC와 MFCC-delta 및 MFCC-acceleration을 모두 중첩한 형태이다. 오디오 신호의 한 프레임은 40ms 단위로 50%의 중첩을 갖고 로그 멜-밴드 에너지 값으로 변환되며, 각 프레임 데이터에 이산 코사인 변환을 적용하여 0차부터 19차까지 MFCC 특징 벡터로 취한다. 또한 MFCC-delta와 MFCC-acceleration은 MFCC 아홉 프레임에 대해 특징을 계산하여, 현재 프레임을 중심으로 시간 경과에 따른 MFCC 계수의 변화를 포함한다. 결과적으로 한 프레임당 위의 세 특징을 합친 60차원의 입력 특징 벡터를 모델에 입력하게 된다. 가우시안 혼합 모델에는 16개 가우시안이 사용되었으며, 각각의 이벤트 클래스에 대해 긍정 및 부정 확률을 추정하는 바이너리 분류 모델이 설정되었다. 여기서 긍정 모델은 해당 이벤트 클래스가 포함된 오디오만을, 부정 모델은 해당 이벤트 클래스가 포함되지 않은 오디오만을 사용하여 학습되었으며, 각 세그먼트당 클래스의 존재 여부는 두 모델 사이의 우도 비에 따라 판단한다.

두 번째로 사용된 비교 알고리즘은 다층 퍼셉트론 모델이다. 다층 퍼셉트론은 뉴런을 수학적으로 모델화한 퍼셉트론으로 하나의 층을 구성하고, 그 층을 다층 구조로 형성하여 네트워크를 구성하는 방법이다. 이 모델은 다양한 비선형 문제를 해결할 수 있는 가장 원시적인 인공 신경망 모델이지만, 앞서 소개된 가우시안 혼합 모델에 비교하여 충분히 높은 인식 성능을 나타낸다고 알려져 있다. 다층 퍼셉트론 모델은 역전파(Backpropagation)를 통해 각 퍼셉트론의 가중치와 바이어스를 학습할 수 있으며, 모델의 학습에는 경사 하강법을 사용하여 예측값과 실제값 사이의 오차를 줄여나간다.

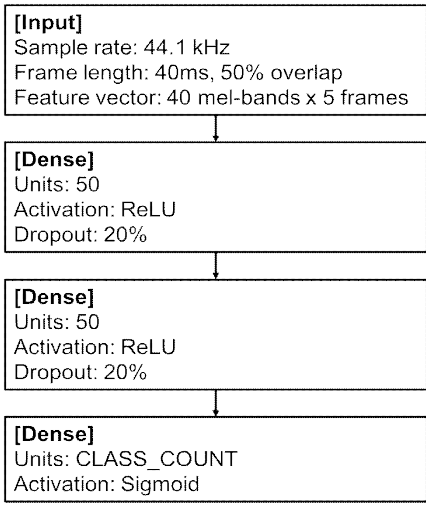

그림 2. 다층 퍼셉트론 모델 블록선도
Fig. 2. Block diagram of multi-layer perceptron model

본 논문에서는 DCASE 2017의 기준 시스템[7]으로 제안된 다층 퍼셉트론 모델을 기반으로 실험을 수행하였다. 한 프레임의 오디오 입력은 0 ~ 22050 Hz를 커버하는 40개의 로그 멜-밴드 에너지를 사용하며, 40ms 단위로 50% 중첩을 통해 생성된다. 이렇게 생성된 특징을 활용해 하나의 입력 특징 벡터는 문맥정보를 고려하기 위해 이를 다섯 프레임 합쳐서 활용한다. 본 연구에서 사용된 다층 퍼셉트론 모델은 그림 2와 같이 50개 유닛을 포함하는 두 개의 은닉층과 이벤트 클래스 개수만큼의 유닛을 포함한 한 개의 출력층으로 구성되어 있다. 각각의 은닉층 뒤에는 과적합(over-

fitting)을 피하고자 20%의 dropout이 적용되었다. 모델은 경사 하강법 기반 최적화를 위한 Adam 알고리즘을 활용하여 학습되었으며, 0.001의 learning rate로 최대 200 epoch 동안 수행된다. 100 epoch의 학습 후, 매 10 epoch을 검토하여 학습의 조기 종료를 적용한다.

IV. 실험 및 결과

1. 실험 방법 및 평가 지표

본 실험에서는 II 장에서 기술한 데이터 세트를 이용하여 음향 이벤트 인식 성능을 검증하였으며 III 장에서 서술한 바와 같이 총 세 가지 알고리즘에 대해 검증 실험을 수행하였다. 본 실험은 DCASE의 음향 이벤트 인식 과제[5,6]의 평가 방식을 따르며, 각각의 실내외 음향 장면에서 발생하는 음향 이벤트로 실내외 각 16종에 대한 인식 결과를 평가한다. 전체 데이터 세트는 실내와 실외에서 획득된 데이터에 대해 각각 4겹 교차 검증(4-fold cross validation)을 통해 획득된 데이터베이스에 대한 음향 이벤트 인식 성능 검증을 수행하였다.

DCASE의 음향 이벤트 인식 과제는 각 인식 모델에 대한 평가 지표(metrics)로 F1-점수(F1-score)와 에러율(error rate)을 제시하고 있다[19]. 각 평가 지표는 1초 단위의 세그먼트 별로 시스템의 출력 결과와 실제값(ground truth)의 대조 결과를 알아야 하며, 이는 그림 3과 같이 네 가지 경우로 분석할 수 있다. 시스템의 출력과 실제값이 같다면 두 값이 모두 참일 경우 true positives(TP), 모두 참이 아닐 경우 true negatives(TN)라 한다. 만약 시스템의 출력과 실제값이 다르다면, 시스템 출력은 참이지만 실제값은 참이 아닐 경우 false positives(FP)라 하고, 반대로 시스템 출력은 참이 아니지만 실제값은 참일 경우 false negatives(FN)라 한다.

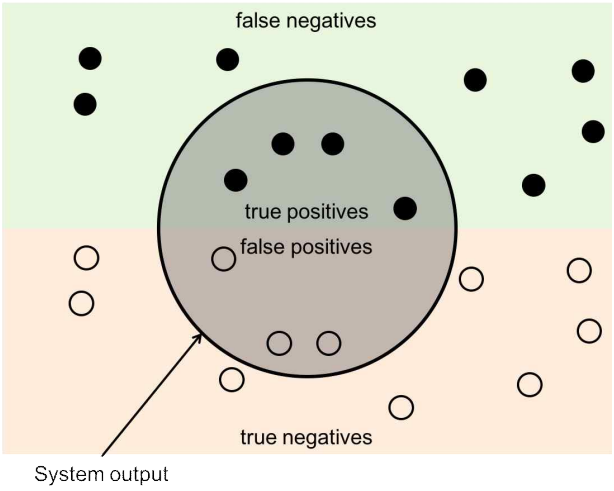

그림 3. Ground truth에 대한 시스템 출력의 시각화
Fig. 3. Visualization of system output to ground truth

F1-점수의 경우 분류 문제를 해결할 때 정확도를 표현하기 위해 사용되는 척도로 식 (1)과 같이 각 세그먼트 k에 대해서 정밀도(precision, P)와 재현율(recall, R)의 조화 평균으로 구할 수 있다. 정밀도란 참인 클래스에 속한다고 출력한 샘플 중 실제로 참인 샘플수의 비율을 의미하며, 재현율이란 실제로 참인 클래스에 속한 샘플 중에 참인 클래스에 속한다고 출력한 샘플 수의 비율을 의미한다.

$$F_1 = \frac{2}{P^{-1} + R^{-1}}, \text{ where}$$
$$P = \frac{\Sigma TP(k)}{\Sigma TP(k) + \Sigma FP(k)}, \quad R = \frac{\Sigma TP(k)}{\Sigma TP(k) + \Sigma FN(k)} \tag{1}$$

에러율은 실제값에 존재하는 전체 이벤트 수에 대한 오판 혹은 오검출 사례를 수치화한 지표로, 모델이 실제값을 잘 예측할수록 낮은 수치를 나타내기에 가장 이상적인 모델의 경우 값이 0이 된다. 다중 음향 이벤트 인식 과제에서 전체 세그먼트에 대한 에러율은 식 (2)와 같이 구할 수 있다. 위 식에서 N(number of reference events)은 실제값에 포함된 세그먼트 별 이벤트의 전체 개수를, S(substitutions)는 시스템이 찾아낸 이벤트가 실제값과 다른 경우, I(insertions)는 FP에서 S를 제외한 경우, 그리고 D(deletions)는 FN에서 S를 제외한 경우를 나타낸다.

$$Error\,rate = \frac{\Sigma S(k) + \Sigma D(k) + \Sigma I(k)}{\Sigma N(k)} \tag{2}$$

2. 실험 결과

본 실험에서는 앞서 기술한 데이터 세트를 이용하여 음

향 이벤트 인식 알고리즘의 성능을 검증하였으며, 이를 위해 두 가지의 실험을 수행하였다. 첫 번째로는 제안하는 음향 이벤트 인식 알고리즘의 컨벌루션 신경망 입력 이미지 구성을 위한 파라미터를 선정하기 위해 몇 개의 시간 프레임을 동시에 하나의 입력으로 구성할 것인지와, 몇 프레임을 건너뛰면서 각 이미지를 획득할 것인지에 대한 파라미터 값을 변경하며 실험을 수행하였다. 두 번째로는 수집된 음향 이벤트 데이터 세트를 이용하여 제안하는 음향 이벤트 인식 알고리즘과 기존 알고리즘의 성능을 비교 검증하였다.

컨벌루션 신경망은 2차원 신호로 변환된 오디오 신호를 입력으로 사용할 경우 음향 신호의 시간적인 특징 및 주파수 특성을 동시에 모델링 할 수 있다. 이때 몇 개의 시간 프레임 문맥(context size)을 하나의 입력 이미지로 구성하는지에 따라서 컨벌루션 신경망으로 입력되는 이미지의 크기가 달라지며, 이에 따라 신경망의 성능도 변할 수 있다. 또한, 컨벌루션 신경망의 입력을 구성할 때 몇 프레임을 건너뛰며(hop size) 이미지를 획득하느냐에 따라 학습 및 평가 셋의 크기가 변하게 되며 이 또한 신경망의 성능에 영향을 줄 수 있다. 따라서 본 실험에서는 상기 기술한 두 개의 파라미터 값을 변경시키며 최적의 값을 찾는 실험을 수행하였고, 실험 결과는 표 4와 같다. 본 실험에서는 수집된 데이터베이스 중 실내 데이터베이스를 이용하여 실험하였으며, 4겹 교차 검증을 통해 F1-점수를 평가 기준으로 하여 수행되었다. 표 4에서 보는 바와 같이, 문맥 크기가 현재 프레임 앞, 뒤로 25개 프레임씩 총 51개 프레임을 사용하고, 홉 사이즈는 10개 프레임씩 건너뛰면서 입력 이미지를 구성할 때 가장 우수한 성능을 보였다. 문맥 크기가 더 커지거나 홉 사이즈가 작아지게 되면 더 우수한 성능을 보일 수도 있으나 본 연구에서는 성능 및 연산량 등을 고려하여 문맥의 크기를 51, 홉 사이즈를 10으로 하여 이후 실험을 진행하였다.

표 4. 문맥 사이즈와 홉 사이즈 변경에 따른 음향 이벤트 인식 결과
Table 4. Acoustic event detection results according to the context and hop size

Context size [frames]	Hop size [frames]			
	5	10	25	50
11	80.1	79.6	77.5	77.5
21	79.9	79.9	79.1	77.5
51	80	80.4	78.4	78.1
101	79.7	80.3	79.1	78.3

음향 이벤트 인식 성능 측정은 II장에서 소개한 음향 이벤트 데이터 세트를 활용하여 진행했으며, 실내외 음향 이벤트 별 상세 인식 결과는 각각 표 5와 6의 내용과 같다. 실내 환경에서 수집된 이벤트 인식 성능은 F1-점수가 80.4%, 에러율이 0.35가 기록된 반면, 주변 환경음의 영향을 더 받는 실외 환경에서 수집된 이벤트 인식 성능은 F1-점수가 73%, 에러율이 0.46이 기록되었다. 각 음향 이벤트 별 인식 결과는 전반적으로 고른 성능을 보이나, 일부 이벤트 클래스에 대해서는 매우 낮은 인식 성능을 보였다. 이는 실내 환경의 Children shouting, Dish cleanup 클래스와 실외 환경의 Speech 클래스의 경우, 수집된 인스턴스 수가 적거나, 음향 이벤트의 발생 시간이 짧아 클래스 불균형(Class imbalance) 문제로 인해 알고리즘이 해당 클래스를 제대로 인식하지 못한 것으로 해석할 수 있다. 또한 실외 환경의

표 5. 실내 환경에서 음향 이벤트 별 상세 인식 결과
Table 5. Detailed results of detection per sound event in an indoor environment

<Indoor>	F1-score [%]	Error Rate	Precision	Recall
Kettle whistle	93.4	0.13	0.928	0.94
Children crying	64.4	0.74	0.62	0.67
Children playing	75.7	0.56	0.666	0.877
Children shouting	0	1	0	0
Dish cleanup	0	1	0	0
Dish rinse sound	81.4	0.38	0.803	0.824
Dishwasher	79.5	0.42	0.77	0.821
Doorbell	91.5	0.18	0.881	0.952
Drawer sound	66.8	0.66	0.672	0.664
Drop impact sound	54.4	0.82	0.611	0.49
Fire alarm sound	91.3	0.18	0.895	0.933
Footfall	78.3	0.43	0.797	0.769
Keyboard sound	92.4	0.16	0.903	0.945
Scream	69.3	0.64	0.668	0.721
Speech	77.2	0.45	0.781	0.763
Water flowing	88.2	0.25	0.851	0.915
Instance-based average	80.4	0.35	-	-

Bird singing 클래스는 어노테이션 된 구간 내에 실제 음향 이벤트 보다 공백 구간이 많이 포함되어 있어 해당 클래스의 음향 특성을 구분해 내는데 실패한 것으로 보인다.

표 6. 실외 환경에서 음향 이벤트 별 상세 인식 결과
Table 6. Detailed results of detection per sound event in an outdoor environment

<Outdoor>	F1-score [%]	Error Rate	Precision	Recall
Bicycle idle horn	72.7	0.57	0.698	0.758
Bird singing	9	0.98	0.612	0.049
Car crash	54.9	0.86	0.574	0.526
Car idle horn	66.3	0.61	0.739	0.602
Car passing	77.3	0.48	0.738	0.813
Car passing horn	54.9	0.7	0.768	0.427
Drop impact sound	68.5	0.61	0.71	0.662
Footfall	74.4	0.52	0.727	0.762
Motorcycle idle horn	68.7	0.66	0.656	0.721
Motorcycle passing horn	65.1	0.63	0.724	0.591
Scream	46.9	0.82	0.664	0.362
Speech	0	1	0	0
Truck idle horn	61.8	0.62	0.808	0.5
Truck passing	61.1	0.71	0.671	0.56
Truck passing horn	45.8	0.82	0.673	0.347
Wind sound	84	0.31	0.857	0.823
Instance-based average	73.0	0.46	-	-

음향 이벤트 인식 성능 비교를 위한 기준 시스템으로는 III장 2절에서 소개한 가우시안 혼합 모델과 다층 퍼셉트론 모델, 그리고 보조 신경망 없이 주 신경망으로만 구성된 CNN 모델이 사용되었다. 가우시안 혼합 모델과 다층 퍼셉트론 모델의 문맥 사이즈와 홉 사이즈는 본 논문에서 제안한 모델과 동일한 입력 파라미터에서 실험을 진행했으나, III장 2절에서 기술한 DCASE의 파라미터에서 가장 높은 성능을 보이는 것으로 확인되었다.

표 7에서 볼 수 있듯이, 제안하는 방법은 실내 음향 이벤트 데이터 세트에서 가우시안 혼합 모델 대비 F1-점수가 19.7%, 에러율이 0.72 만큼 우수한 성능을 보였으며, 다층 퍼셉트론 모델 대비 F1-점수가 3.0%, 에러율이 0.01 만큼 우수한 성능을 보였다. 또한, 실외 음향 이벤트 데이터 세트에 대해서도 가우시안 혼합 모델 대비 F1-점수가 7.2%, 에러율이 0.16 만큼 우수한 성능을 보였으며, 다층 퍼셉트론 모델 대비 F1-점수가 3.0%, 에러율이 0.01 만큼 우수한 성능을 보였다. 마지막으로 보조 신경망을 활용한 이중 CNN 알고리즘의 유효성을 검증하기 위해, 주 신경망으로만 구성된 CNN 모델(w/o aux)과 보조 신경망을 적용한 이중 CNN 모델(w/ aux)의 성능을 비교하였다. 그 결과, 제안하는 이중 CNN 모델이 모든 평가 척도에서 성능 개선을 보였으며, 이를 통해 프레임 별 이벤트 존재 여부를 확인하는 것이 이벤트 인식 성능 개선에 도움을 주는 것으로 확인되었다.

표 7. 음향 이벤트 인식 시스템 성능 측정 결과
Table 7. Performance test for sound event detection systems

	Indoor		Outdoor	
	F1-score [%]	Error Rate	F1-score [%]	Error Rate
GMM	60.7	1.07	65.8	0.62
MLP	77.4	0.36	70.0	0.47
CNN (w/o aux.)	79.2	0.36	72.1	0.47
Proposed (w/ aux.)	80.4	0.35	73.0	0.46

V. 결론 및 향후 연구 방향

본 논문에서는 음향 이벤트 인식 기술 개발에 필요한 데이터 세트의 설계 및 제작, 그리고 이중 CNN 구조를 특징으로 하는 음향 이벤트 인식 알고리즘을 제안하였다. 먼저, 데이터 세트는 DCASE 챌린지를 위해 제작된 TUT 데이터베이스의 상세 규격에 준하여 제작하였으며, 실내와 실외 환경에서 발생할 수 있는 음향 이벤트 클래스를 먼저 정의한 뒤 수집을 진행하였다. 실내외에서 수집한 데이터 세트의 규모는 스테레오와 바이노럴 녹음 각각에 대해 13시간 9분 분량의 파일로, 클래스당 평균 254개 음향 이벤트를 포함한다. 이는 DCASE 2017 챌린지가 1시간 31분 분량의 데이터 세트를 제공한 것에 비해 더 큰 규모이며, 각 클래스당 평균 1.7배 정도 더 많은 인스턴스 수를 갖고 있다.

다음으로 이벤트 존재 여부 오판에 대해 강인한 이중 CNN 기반 음향 이벤트 인식 시스템을 제안하고, 성능 평가를 위한 실험을 진행하였다. 제안된 음향 이벤트 인식 시스

템은 컨벌루션 층으로 구성된 주 인식 네트워크와 이벤트 존재 여부에 대한 오검출을 줄이기 위해 프레임별 이벤트 존재 여부를 0과 1로 판단하는 보조 네트워크가 적용된 이중 CNN 구조로 되어 있다. 알고리즘의 성능 평가는 DCASE 챌린지의 음향 이벤트 인식 과제와 같은 방식으로, II 장에서 소개된 데이터 세트를 활용하여 진행하였다. 본 실험에 앞서 컨벌루션 신경망은 입력되는 이미지의 크기에 따라 성능에 영향을 줄 수 있으므로 문맥의 크기와 홉 사이즈에 따른 성능 분석 실험을 진행하였으며, 각각 51과 10으로 설정하였을 때 가장 높은 성능을 보였다. 성능 평가를 위해 진행된 실험에는 DCASE 챌린지의 2016년 기준 시스템인 가우시안 혼합 모델과 2017년 기준 시스템인 다층 퍼셉트론 모델이 대조군으로 활용되었다. 실험 결과, 제안된 음향 이벤트 인식 알고리즘은 F1-점수 및 에러율에서 모든 기준 시스템보다 높은 성능을 보였다.

향후 연구로는 인스턴스 수가 부족한 음향 이벤트는 추가 수집을 진행하고, 실내외에서 발생할 수 있는 위험 또는 비위험 음향 이벤트 클래스를 추가 정의하여 데이터 세트의 규모를 늘려나갈 예정이다. 또한, 실내외로 구분된 현재의 음향 장면을 세분화하여 추가 수집을 진행하고, 메타데이터에 음향 장면 항목을 추가하여 어노테이션 작업을 진행할 예정에 있다. 이 밖에도, 각 이벤트 클래스의 음향 특성에 따라 인식 성능이 크게 떨어지는 상황을 극복할 수 있는 새로운 구조의 인식 알고리즘에 대한 연구도 병행하여 진행할 예정이다.

참 고 문 헌 (References)

[1] A. Temko et al., “CLEAR evaluation of acoustic event detection and classification systems,” Lecture Notes in Computer Science, vol.4122, pp.311-322, 2007.

[2] D. Stowell et al., “Detection and classification of acoustic scenes and events,” IEEE Transactions on Multimedia, vol.17, no.10, pp.1733-1746, 2015.

[3] DCASE Community, http://dcase.community/community_info

[4] J. Portêlo et al., “Non-Speech Audio Event Detection,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2009.

[5] DCASE 2016 Task3 Sound event detection in real life audio, http://www.cs.tut.fi/sgn/arg/dcase2016/task-sound-event-detection-in-real-life-audio

[6] DCASE 2017 Task3 Sound event detection in real life audio, http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/task-sound-event-detection-in-real-life-audio

[7] A. Mesaros et al., “DCASE 2017 Challenge Setup: Tasks, Datasets and Baseline System,” Detection and Classification of Acoustic Scenes and Events (DCASE), 2017.

[8] A. Mesaros, T. Heittola, and T. Virtanen, “TUT Database for Acoustic Scene Classification and Sound Event Detection,” 24th European Signal Processing Conference (EUSIPCO), pp. 1128-1132, 2016.

[9] S. Adavanne, G. Parascandolo, P. Pertila, T. Heittola, and T. Virtanen, “Sound event detection in multichannel audio using spatial and harmonic features,” Detection and Classification of Acoustic Scenes and Events (DCASE), 2016.

[10] I. Jeong, S. Lee, Y. Han, and K. Lee, “Audio event detection using multiple-input convolutional neural network,” Detection and Classification of Acoustic Scenes and Events (DCASE), 2017.

[11] S. Adavanne, and T. Virtanen, “A report on sound event detection with different binaural features,” Detection and Classification of Acoustic Scenes and Events (DCASE), 2017.

[12] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” IEEE Conference on Computer Vision and Patter Recognition (CVPR), 2016.

[13] Large Scale Visual Recognition Challenge (LSVRC), http://image-net.org/challenges/LSVRC/ImageNet, http://www.image-net.org/

[14] ImageNet, http://www.image-net.org/

[15] Y. Jung, S. Seo, W. Lim, and H. Kim, “Design and construction of Acoustic Database for developing Sound Event Detection technique,” IEIE Summer General Conference, June, 2018

[16] D. P. Kingma, and J. Ba, “Adam: A method for stochastic optimization,” Proceedings of the 3rd International Conference on Learning Representations (ICLR), 2014.

[17] TensorFlow, https://www.tensorflow.org/

[18] Keras, https://keras.io/

[19] Metrics For sound event detection tasks, http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/metrics

저 자 소 개

서 상 원

- 2015년 : 한국과학기술원 문화기술전공 공학석사
- 2015년 ~ 현재 : ETRI 실감AV연구그룹 연구원
- ORCID : https://orcid.org/0000-0002-4286-6537
- 주관심분야 : 오디오 신호처리, 기계학습

임 우 택

- 2010년 : 광운대학교 전자공학과 공학사
- 2012년 : 광운대학교 전자공학과 공학석사
- 2012년 ~ 현재 : ETRI 실감AV연구그룹 연구원
- 주관심분야 : 오디오 신호처리, 기계학습

정 영 호

- 1992년 : 전북대학교 전자공학과 공학사
- 1994년 : 전북대학교 전자공학과 공학석사
- 2006년 : 충남대학교 전자공학과 공학박사
- 2011년 ~ 2017년 : 과학기술연합대학원대학교(UST) 이동통신및디지털방송공학과 겸임교수
- 1994년 ~ 현재 : ETRI 실감AV연구그룹 책임연구원
- ORCID : https://orcid.org/0000-0001-9552-8593
- 주관심분야 : 음향인식, 머신러닝, 오디오 신호처리

이 태 진

- 2014년 : 충남대학교 전자전파정보통신공학과 공학박사
- 2002년 ~ 2003년 : 일본 Tokyo Denki University, 방문연구원
- 2000년 ~ 현재 : ETRI 실감AV연구그룹 책임연구원/전문위원
- 주관심분야 : 오디오 부호화, 실감음향, 오디오 신호처리

김 휘 용

- 1994년 : 한국과학기술원 전기및전자공학과 공학사
- 1998년 : 한국과학기술원 전기및전자공학과 공학석사
- 2004년 : 한국과학기술원 전기및전자공학과 공학박사
- 2003년 ~ 2005년 : ㈜애드팍테크놀러지 기술연구소 멀티미디어팀장
- 2006년 ~ 2010년 : 과학기술연학대학원대학교(UST) 이동통신및디지털방송공학과 겸임교수
- 2013년 ~ 2014년 : Univ. Southern California(USC) 멀티미디어통신연구실 방문연구원
- 2005년 ~ 현재 : ETRI 실감AV연구그룹장
- ORCID : https://orcid.org/0000-0001-7308-133X
- 주관심분야 : 비디오/오디오 신호처리 및 부호화, 컴퓨터 비전, UHD/3D/HDR/VR 등 실감미디어서비스