Journal of Korea Robotics Society (2022) 17(3):273-281
https://doi.org/10.7746/jkros.2022.17.3.273 ISSN: 1975-6291 / eISSN: 2287-3961

Random Forest를 결정로직으로 활용한 로봇의 실시간 음향인식 시스템 개발

A Real-Time Sound Recognition System with a Decision Logic of Random Forest for Robots

송 주 만[1] · 김 창 민[1] · 김 민 욱[1] · 박 용 진[1] · 이 서 영[1] · 손 정 관†
Ju-man Song[1], Changmin Kim[1], Minook Kim[1], Yongjin Park[1],
Seoyoung Lee[1], Jungkwan Son†

Abstract: In this paper, we propose a robot sound recognition system that detects various sound events. The proposed system is designed to detect various sound events in real-time by using a microphone on a robot. To get real-time performance, we use a VGG11 model which includes several convolutional neural networks with real-time normalization scheme. The VGG11 model is trained on augmented DB through 24 kinds of various environments (12 reverberation times and 2 signal to noise ratios). Additionally, based on random forest algorithm, a decision logic is also designed to generate event signals for robot applications. This logic can be used for specific classes of acoustic events with better performance than just using outputs of network model. With some experimental results, the performance of proposed sound recognition system is shown on real-time device for robots.

Keywords: Sound Event Detection, Deep Learning, Robot Implementation, Audio Signal Processing, Machine Learning, Real-Time Implementation

1. 서 론

최근에는 다양한 환경에서 사용할 수 있도록 여러 가지 로봇들이 개발되고 있다. 이러한 로봇들은 주어진 작업을 수행하기 위해 다양한 센서를 사용하여 복잡한 현실 세계를 감지하고 인간의 신체 능력을 모방한다. 로봇의 다양한 센서 중 마이크는 소리를 듣고 그 신호를 분석하여 의미 있는 정보를 얻는 귀의 능력을 모방하기 위해 사용되어왔다. 마이크에서 중요한 정보를 얻기 위해 일부 연구자들은 다양한 소리 이벤트를 감지할 수 있는 딥 러닝 기술을 사용해 왔다. Visual Geometry Group이라는 이름의 연구 그룹은 다양한 크기의 VGG model들을 제안했고[1] Hershey 등은 VGG model들과 유사한 모델링을 통해 VGGish 모델을 제안하였다[2]. ETRI는 음향 분류를 위해 다른 convolutional neural network (CNN) 변형을 제안하였으며[3], LG전자 로봇선행연구소에서는 앙상블 기법과 함께 여러 VGGish 모델을 사용하였다[4]. 오프라인 시스템에서 실제 오디오 파일의 사운드 이벤트를 감지하기 위해 Random Forest (RF) 란[5] 일반적인 분류기 중 하나를 변형한 deep RF가 도입되기도 하였다[6]. 이러한 연구들은 네트워크 모델을 구성할 때, fully connected 레이어에 average pooling 레이어, softmax 레이어 또는 임계값을 결정 로직으로 사용하였다.

여러 논문들은 비교적 제한적인 환경에서의 음향 이벤트 감지 성능을 보여주었지만, 실시간 환경에서는 매우 다양한 음향 환경이 있을 수 있다. 첫째로, 분류 대상에 속하지 않은 다른 음향 이벤트가 너무 많아서 실시간 환경의 SNR (Signal to Noise Ratio)은 다양하게 변경될 수 있다. 특히, 로봇에서 음향 인식 시스템을 사용하기 위해서는 로봇에는 다양한 환경에서 작동하는 모터들이 있다. 그런 다양한 잡음들을 극복하기 위해 강인한 음향 인식 시스템을 구현하려는 많은 연구들이 있었다[7,8].

Received : May. 27. 2022; Revised : Jun. 29. 2022; Accepted : Jul. 6. 2022
※ This project was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (No. 2020-0-00857)
1. Researcher, LG Electronics, Seoul, Korea (juman.song, changmin.kim, minook83.kim, yongjinn.park, seoyoung28.lee@lge.com)
† Researcher, Corresponding author: LG Electronics, Seoul, Korea (jungkwan.son@lge.com)

두 번째로, 음향 인식 시스템의 실시간 구현을 위해서는 잔향 정보가 고려되어야 한다. 음향 이벤트의 소리는 장소에 따라 다양한 방식으로 전파되며, 음원의 움직임이 있으면 잔향 특성이 보다 시시각각 다양하게 변화된다. 따라서 이러한 음향 인식 시스템과 유사한 문제를 해결하기 위해 음성 인식 시스템에서는 생성적 적대 네트워크를 사용했다[9]. 로봇의 경우 로봇이 움직이면, 로봇에 장착된 마이크 또한 움직이기 때문에, 음향 경로의 복잡성이 더 커지게 된다. 이처럼 복잡하게 변화되는 잔향 특성이 모델의 학습 환경의 잔향 특성과 차이가 커질수록 음향 인식 네트워크 모델의 성능이 저하된다는 것 또한 연구되어 왔다[10].

마지막으로, 음향 인식 시스템은 로봇과 같은 실시간 시스템에서 작동해야 한다. 로봇의 많은 어플리케이션들에서 감지된 이벤트 결과를 사용하려면 빠른 응답 시간이 필요하다. 이를 위해, 다양한 로봇들이 NVIDIA Jetson AGX Xavier[11]와 같은 장치에서 실시간으로 딥러닝 기술을 사용하고 있다. 본 논문에서도 NVIDIA Jetson AGX Xavier 개발 키트를 사용하여 실시간 음향 인식 시스템을 구축하였으며, 모델 추론 성능은 TensorRT[12]를 이용하여 최적화하였다.

본 논문은 VGG11 모델을 사용하여 실시간으로 추론을 수행하는 음향 인식 시스템을 제안한다. 잡음이 많은 입력에 대한 강인함을 얻기 위해 다양한 잔향 및 SNR 환경에서 데이터 증강이 수행되었으며, 입력 버퍼의 실시간 정규화 방법도 설계하였다. 여러 로봇 어플리케이션들의 요구를 반영하기 위해 RF를 이용한 결정 로직을 설계하여 음향 이벤트 감지 메시지를 보내도록 설계하였다.

2. 음향인식 모델 설계

본 섹션에서는 로봇에서 실시간으로 추론을 할 수 있는 네트워크 모델을 제안한다. 우선, 서로 겹치지 않고 명확하게 정의된 다양한 클래스들을 선택한다. 그 후 VGG11 모델을 실시간 프로세스들에서 사용하도록 설계하였다. 또한, 복잡한 음향 환경에 대응하기 위해 다양한 임펄스 응답과 2가지 SNR값으로 믹싱된 음향 DB들로 모델 학습을 수행한다.

2.1 다양한 음향 클래스 선정

본 논문에서는 가정 환경에서 나타날 수 있는 다양한 음향 이벤트들을 서로 겹치지 않으며 명확한 정의가 되도록 클래스들을 선정하였다. 만약, 일부 클래스들이 서로 겹치는 부분이 있을 경우 개발된 네트워크는 해당 클래스에 대한 성능이 떨어질 수 있다. 예를 들어 'children playing' 클래스가 'speak' 클래스와 같이 클래스 리스트에 들어가게 되면, 어린이들이 노는 환경에 여성의 목소리가 들어 있는 경우가 많아 학습된 모델은 'children playing' 클래스의 소리를 여성의 목소리가 포함된 'speak'로 인식할 수 도 있다. 클래스들간의 배타적 정의가 중요한 요소인 것처럼, 이벤트의 명확성은 음향 인식 시스템에 있어서도 중요하다. 특정 클래스의 정의가 모호한 경우 수집된 음향 파일이 훈련된 네트워크 모델 학습에 부정적인 영향을 미칠 수 있다. 이와 같은 것들을 고려하여 본 논문에서는 로봇에서 인식할 29개의 클래스를 선택하였다.

29개의 선정된 클래스 외에 'noise'라는 이름의 추가 클래스를 하나 더 선정하였다. 해당 클래스는 입력된 오디오 버퍼에 29개의 클래스 중 선택할 클래스가 없을 때 선택되어야 할 필수적인 클래스이다. 본 클래스의 DB는 사람이 없고 특별한 이벤트가 없는 무인환경의 거실에서 소리를 녹음하여 수집하였다. 다른 클래스는 다양한 방법으로 수집하였다. 일부 음원은 일반적인 가정 환경이나 실험실에서 녹음하였다. 또 다른 음원들은 크라우드 워커를 사용하여 취득되기도 하였다. 선정된 클래스 목록은 [Table 1]에 정리되어 있다.

2.2 데이터 증강

제안된 네트워크 모델을 로봇에 적용해서 다양한 환경에 강인하게 음향인식을 사용하기 위해 다양한 방법으로 데이터를 증가하였다. 첫 번째 방법은 다양한 임펄스 응답에 대한 데이터 증강 방법이다. 임펄스 응답은 [8]과 같이 학습에 사용되

[Table 1] A List of Primary Sound Events

Class Number	0	1	2	3	4	5	6	7
Class Name	applause	baby cry	bike bell	blender	cat	dog	doorbell	doorlock
Class Number	8	9	10	11	12	13	14	15
Class Name	drilling	electric shaver	engine idling	fan motor	glass break	gun shot	jackhammer	keyboard
Class Number	16	17	18	19	20	21	22	23
Class Name	kitchen hood	klaxon	knock	laugh	motorcycle	music	noise	scream
Class Number	24	25	26	27	28	29		
Class Name	siren	sneeze	snore	speak	toilet	vibration		

는 환경과 테스트에 사용되는 환경이 달라지면 성능이 하락하기 때문에 실시간 음향 인식에 있어서 중요한 요소이다. 이에 따라, 다양한 임펄스 응답을 얻기 위해 여러 공간에서 측정을 하였고 0.21에서 1.10 사이의 12개의 각각 다른 T_{60}를 가지는 임펄스 응답을 얻었다. 추가적으로 SNR에 따라 음향 인식 성능이 크게 변할 수 있으므로, 측정된 임펄스 응답들을 원본 파형 오디오에 필터링하며 10 dB 및 20 dB의 두 가지 SNR환경으로 카페에서 취득된 잡음을 추가하여 데이터 증강을 하였다.

잡음 취득은 서울의 한 카페에서 1시간 동안 녹음되었으며, 임의의 시작점을 사용하여 신호의 무작위성을 보장하였다. 최종적으로 12개의 임펄스 응답과 2가지 SNR조합인 24가지 데이터 세트에 원본 세트까지 더해 총 25가지의 증강된 데이터 세트를 사용하였다.

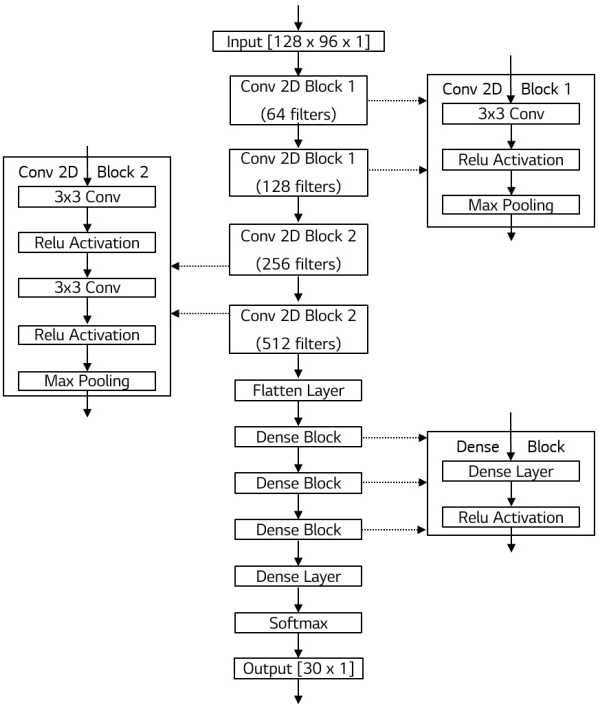
2.3 음향 인식 Feature 추출

본 논문에서 사용되는 증강된 DB의 sample rate는 16 kHz, 즉 1초에 16000샘플을 가지고 있다. 이 때, 한 프레임을 0.02초인 320개의 샘플로 구성하였고 길이 640 샘플의 Hann window로 모든 320개 샘플을 이동하여 프레임별로 파워 스펙트럼으로 변환했다. 퓨리에 변환의 경우 1024포인트를 사용하였으며, 1.92초 오디오 파일에서 총 96프레임의 스펙트럼을 추출하고, 최종적으로 각 스펙트럼을 멜스케일의 128 bin으로 병합하였다. 이에 따라, 모델에 입력되는 feature의 최종 크기는 [128 x 96 x 1]이 된다. 해당 feature는 실시간 테스트를 위한 로봇에 동일하게 적용된다.

2.4 음향 인식 모델 설계

제안된 음향 인식 시스템을 실시간 로봇에서 사용하기 위해 본 논문에서는 VGG 모델들[1] 중 VGG11 모델을 사용하였다. VGG모델은 음향인식 분야에서 일반적으로 사용되는 모델로 RF를 이용한 결정로직의 효과를 증명하고자 본 논문에서도 사용되었다. 다만, 실시간 동작성을 위해 모델 크기를 줄이고자 CNN 레이어에서 padding을 적용하지 않았다. Padding을 적용하지 않으면 CNN 레이어를 통과할 때마다 레이어의 출력이 입력보다 작아지게 되므로, 너무 많은 CNN 레이어를 사용하면 클래스 개수보다 작은 CNN block의 출력이 나오게 된다. 때문에, padding을 사용하지 않는 경우에 VGG모델들 중에서 사용 가능한 모델은 VGG11이었다. 때문에, 본 논문에서는 VGG11모델을 이용해 학습을 진행하였다.

각 CNN block은 [3 x 3] 사이즈의 커널을 사용한 CNN 레이어, ReLU 활성화 레이어를 [Fig. 1]과 같이 가지고 있다. 또한,



[Fig. 1] The block diagram of sound recognition model VGG11

conv block 1은 CNN 레이어를 1개 conv block 2는 CNN 레이어를 2개 가지고 있다. 커널은 He 정규 분포[13]로 초기화하였으며, 각각의 conv block은 64, 128, 256, 512 개의 필터를 사용하였다. Conv block들을 거친 이후에 flatten 레이어와 3개의 dense block을 지나 추가적인 dense 레이어를 거친 후 softmax 레이어를 통해 1 x 30 의 음향 인식 확률 벡터를 만든다. Dense block은 1개의 dense 레이어와 1개의 ReLU 활성화 레이어로 구성되어 있다. VGG11 모델의 전체 구조는 [Fig. 1]에 정리되어 있다.

3. 로봇에서의 음향 인식 실시간 추론

3.1 실시간 추론을 위한 추론 주기 결정 방법

로봇에서 음향 인식 시스템을 사용할 때 가장 어려운 부분은 얼마나 자주 추론을 수행할지를 결정하는 것이다. 음향 이벤트가 발생했을 때, 빠른 응답속도를 얻기 위해서는 모델의 입력이 되는 feature의 한 프레임이 업데이트 될 때 마다 추론을 수행하는 것이 좋다. 하지만, 로봇에는 여러 가지 어플리케이션들이 동시에 동작하는 환경이 대부분이며, 이에 따라, 음향인식에 할당되는 리소스는 제한이 많이 된다. 제한된 환경에서 안정적으로 음향인식 모델을 추론하기 위해서는 안정적이면서 실시간으로 추론을 할 수 있는 추론 주기를 환경에 맞

춰 결정해줘야 한다. 이를 위해서는 한 번의 음향인식 추론을 수행하는데 걸리는 최대 시간과 한 번의 추론에 필요한 실시간 오디오 샘플의 개수를 함께 고려해야 한다.

모든 실시간 오디오 샘플들에 대하여 추론을 수행하기 위해서는 한 번의 추론에 걸리는 시간보다 더 많은 시간에 해당되는 오디오 샘플들이 업데이트된 오디오 버퍼와 그에 해당하는 feature를 이용하여 한 번의 추론이 수행되어야 한다. 이 때, 한 번의 추론에 사용된 새롭게 업데이트 되는 feature 프레임의 개수를 본 논문에서는 추론 주기 s라고 정의하였다. 또한, 한 프레임의 시간을 t_{frame}, i번째 추론에 걸린 시간을 t_i이라 하였을 때, 한 프레임의 시간을 t_{frame}이라 하면 다음과 같은 부등식이 성립한다.

$$\max(t_i) < t_{frame} \times s, where\ i > 0 \tag{1}$$

$$s > \max(t_i)/t_{frame,} \tag{2}$$

위의 부등식에서 $\max(t_i)$ 및 t_{frame}는 각각 1보다 작다고 가정되었다. 식 (2)가 충족되면 오디오 샘플이 누락되지 않고 강인한 음향 인식 결과를 얻을 수 있다. 따라서 로봇에서 추론을 반복 수행하여 최대 추론 시간 $\max(t_i)$을 확인하여 식 (2)를 통해 최소 추론 주기 s를 얻을 수 있다. 하지만, 추론에서의 f1-score는 추론 주기 s가 변경됨에 따라 같이 변경될 수도 있다. 정확한 추론을 위해서는 해당 버퍼에 충분한 음향 이벤트 샘플이 있어야 하며, 어떤 음향 이벤트를 대상으로 하느냐에 따라 충분한 샘플의 수 도 변경될 수 있다. 따라서, 최적의 추론 주기 s 는 경험적으로 결정될 수 있다.

3.2 실시간 입력 정규화 방법

실시간 음향 인식 시스템에서 고려되어야 할 또 하나의 중요 포인트는 음향 신호의 크기 변화에 대한 통제의 문제가 있다. 고양이 울음 소리, 강아지 짖는 소리와 같은 음향 이벤트들은 음원이 이동하며 발생할 수 있기 때문에, 입력 신호의 진폭이 실시간으로 변화될 수 있다. 또한, 음향을 취득하는 로봇도 이동할 수 있기 때문에, 이러한 모든 진폭변화를 데이터 증강으로 모델 학습에 반영시키기는 어렵다. 따라서 본 논문에서는 멜스펙트로그램의 최대값을 확인하여 실시간 입력 정규화를 각 음향인식 모델의 추론에 사용하였다.

[14]에서는 멜스펙트로그램 $Mel(f)$가 가중 평균 스펙트로그램의 결과로 아래와 같이 정의될 수 있다고 한다.

$$Mel(f) = \Sigma_j |S(f)|^2 \cdot \Gamma(j), \tag{3}$$

$S(f)$는 스펙트로그램을 $\Gamma(j)$는 j번째 멜필터를 뜻한다. 식 (3)은 결국 아래의 식과 같이 멜스펙트로그램에 관한 비례식으로 표현될 수 있다.

$$Mel(f) \propto |S(f)|^2 \propto |f|^2 \propto \max|\boldsymbol{\chi}|^2, \tag{4}$$

χ는 입력 오디오 벡터를 뜻한다. 따라서, 한 번의 추론을 수행할 때, feature 프레임을 구성하는 최대값을 저장해 놓으면, 각각의 추론 직전에 오디오 샘플을 정규화할 수 있다. 현재 프레임 이 n번째 프레임 일때, feature 한 개를 구성하는 프레임들의 각각의 최대값들로 구성된 벡터를 $m(n)$은 아래와 같이 정의할 수 있다.

$$\boldsymbol{m}(n) = [m(n), m(n-1), \ldots, m(n-N+1)]^T \in R^{N\times 1}, \tag{5}$$

$m(n)$은 n번째 프레임의 절대값 중 최대값을, N은 한 번의 추론에 사용되는 프레임 개수를 말한다. 그렇다면, 멜스펙트로그램의 실시간 표준화는 아래와 같이 수행될 수 있다.

$$\widehat{Mel(f)} = \frac{Mel(f)}{(\max(m)+\epsilon)^{2'}} \tag{6}$$

ϵ은 로봇 소프트웨어에서 사용 가능한 가장 작은 숫자로 분모가 0이되어 멜스펙트로그램 값이 무한대로 발산되는 것을 방지하기 위한 값이다.

4. Random Forest를 이용한 결정 로직

위와 같은 과정을 통해 각각의 추론 주기의 프레임마다 음향 인식 확률 벡터를 얻을 수 있게 되었다. 하지만, 확률 벡터만으로는 로봇의 응용 프로그램에 인식 결과를 언제 보내는가가 여전히 문제이다. 따라서 이번 섹션에서는 해당 문제를 해결하기 위해 손 글씨 숫자 인식과 같이 간단한 분류기로 사용되어 온 RF를 이용한 결정 로직을 소개한다.

4.1 이벤트 결정 로직

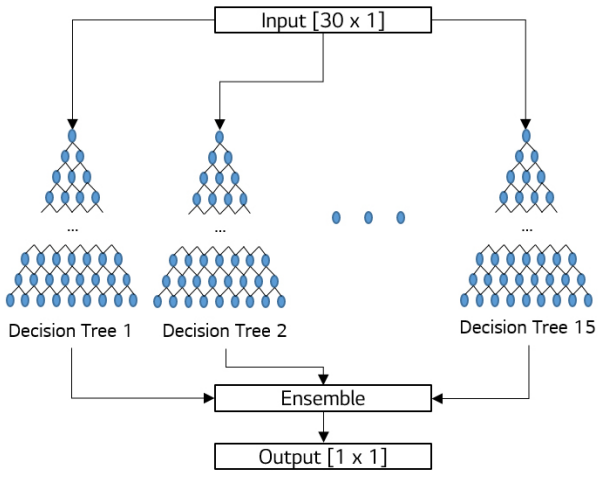
음향 인식 확률 벡터에서 감지된 이벤트를 결정하는 가장 쉬운 방법은 매번 가장 높은 확률 값의 이벤트를 취하는 것이다. 그러나, [Table 1]의 대상 클래스에 포함되지 않은 음향 이벤트, 예를 들어 세탁기 소리가 발생한다면, 주방 후드 클래스처럼 세탁기 소리와 유사한 소리의 확률 값이 가장 높을 수 있다. 이런 경우 정상적인 결정 로직의 결과는 'noise' 클래스여야 하기 때문에, 세탁기 소리로 결정함은 잘못된 결과를 얻게

되는 것이다. 또한 모델의 추론 결과는 음원 또는 로봇의 구동 소음이 유입되거나 불확실한 소음이 발생됨에 따라 변경될 수도 있다. 따라서 정확한 음향 인식 결과 도출을 위해서는 다양한 확률 벡터로부터 인식된 이벤트를 결정하기 위한 결정 로직이 필요하다.

로봇에서 음향인식을 사용하기 위해 필요한 결정로직을 실험적으로 결정하고자 여러 방식의 튜닝을 진행했으며, 실험의 결과에 따라 결정 로직을 설계하는 방식이 결과적으로 각각의 이벤트별 확률값을 조건으로 결정하는 방식으로 이어졌고, 그와 같은 과정을 데이터 학습을 통해 얻을 수 있는 RF를 본 논문에서는 최종 결정 로직으로 선정하였다.

RF를 결정 로직으로 사용하기 위해 모델의 추론 결과인 확률 벡터를 RF에 입력으로 전달한다. 그러면 RF는 어떤 이벤트를 선택해야 하는지 결정하고 선택된 클래스의 인덱스인 정수를 출력한다. 이 RF 로직은 모델의 출력인 확률 벡터와 실제 클래스 번호로 학습된다. 'noise' 클래스 또한 RF의 'noise' 클래스를 학습하는 데에도 사용된다. RF는 [Fig. 2]에 표시된 대로 최대 깊이가 100인 15개의 decision tree로 구성된다. RF의 깊이 및 decision tree 개수는 여러 번의 테스트에 따라 가장 좋은 성능을 보인 개수로 결정하였다.

RF는 보다 구체적인 타겟 이벤트를 인식하는 데도 사용할 수 있다. 섹션 2.1 에서 설명한 것처럼 [Table 1]에는 가능한 많은 클래스인 30개의 클래스들이 있다. 그러나 일반적으로 하나의 응용 프로그램들은 모든 클래스들을 필요로 하지는 않는다. 보통 실시간 어플리케이션들은 소수의 클래스만 필요로 한다. 본 논문에서는 타겟을 가정용 로봇으로 설정하여 RF 로직의 타겟 이벤트를 [Table 2]에 명시한 9개의 클래스로 축소시켰다. 이와 같은 상황에서, 9개의 클래스에 포함되지 않은 클래스들은 'noise' 클래스로 인식 결과가 변경되어야 한다.



[Fig. 2] The block diagram of random forest logic

예를 들면, 인식 결과가 'fan' 클래스일지라도 실제 발생한 음향 이벤트는 로봇에서 멀리 떨어진 지점에서 발생한 'kitchen hood'인 경우도 있으며, 이렇게 모델의 결과가 잘못되었을 지라도, RF에서는 실제 발생한 이벤트로 결정하여야 한다. 이런 경우에는 'kitchen hood' 소리의 인식 확률 벡터를 RF에서 학습시켜 해당 음향 확률 벡터의 패턴이 RF를 통해 학습되어 음향 인식 결과를 개선할 수 있다. 이는 기존의 임계값을 이용한 결정로직보다 더 복잡한 판단을 내릴 수 있게 됨으로써 보다 높은 성능을 얻을 수 있다. 또한, 일부 'fan' 소리는 'kitchen hood'로 감지될 수 있지만 해당 음향의 인식 결과를 'noise'로 처리해 줌으로써도 최종 성능을 높일 수도 있다.

RF 결정 로직은 sklearn-porter[15]라는 파이썬 툴킷을 사용하여 실시간 시스템으로 로봇에 적용할 수 있다. 해당 툴킷은 파이썬의 RF 모델로부터 C++ 코드를 자동으로 생성할 수 있게 해줌으로써 C++ 언어로 설계된 실시간 디바이스 프로그램에 쉽게 적용할 수 있다. 자동 생성된 코드는 매우 빠르게 동작할 수 있어 실시간 동작성에 문제가 없다.

5. 실험 결과

5.1 VGG11 모델 학습

모델 학습에 사용되는 데이터는 최대 10초 길이의 음향 파일들로 구성되어 있다. 모든 오디오는 16 kHz로 resampling되었으며 최대값으로 정규화를 거쳤다. 하나의 클래스는 약 1시간 정도 분량이 있지만, 일부 클래스들은 1시간 미만의 분량을 가지고 있다. 모델의 feature들은 30 클래스 데이터 세트에서 추출되었다. 학습된 모델을 평가하기 위해 81%의 데이터를 학습에 사용하고 9%의 데이터를 검증에 사용하였다. 나머지 데이터 세트의 10%는 테스트에 사용되었다.

제안된 네트워크는 기본 파라미터의 Adam optimizer[16]를 사용하여 최적의 모델을 학습하였다. Categorical focal loss 함수를 $\alpha = 0.5$, $\gamma = 2$로 설정하여 사용하였다. 효율적 학습을 위해 180 크기의 batch를 사용하고 최대 epoch는 150으로 설정했다. 2개의 NVIDIA 2080Ti GPU를 사용하여 증강된 데이터 세트의 VGG11 모델을 훈련하는 데에는 약 4주가 걸렸다. 이

[Table 2] A List of Random Forest Sound Events

# of Class	0	1	2	3	4	5	6	7	8
Class Name	cat	dog	doorbell	doorlock	glass break	kitchen hood	noise	scream	speak

는 클래스, feature 또는 다른 파라미터들을 변경하는 데 너무 오래 걸리는 시간이다. 만약 더 많은 SNR환경도 학습에 사용되면 4주 이상의 시간이 소요된다. 하지만, RF 결정 로직은 사전에 훈련된 모델의 음향 인식 확률 벡터로 훈련하여 훨씬 짧은 시간이 소요된다.

5.2 Random Forest 결정 로직 학습

RF 결정 로직 훈련에는 이벤트가 없는 무음구간을 'noise' 클래스로 학습시키기 위해 이벤트 사이 무음구간도 존재하는 stream 음원을 만들어 사용하였다. 해당 stream 음원에 대한 설명은 5.4에 작성되어 있다. 생성된 음원은 실시간 동작성 평가를 위해 버퍼 핸들 방식을 통하여 제안된 음향 인식 시스템에 전송된다. 버퍼 핸들 방식은 256 샘플의 오디오 데이터를 음향 인식 시스템으로 보내고 시스템 내부 버퍼는 320 샘플의 프레임을 feature 추출 시스템으로 전달된다. 그 후 모든 추론 주기 s마다 모델에서 음향 인식 확률 벡터를 추론한다. 추론된 확률 벡터에서 독립적인 이벤트들 중 80%에 해당하는 이벤트들의 확률 벡터는 RF 학습에 사용하였고 나머지 20%는 RF 테스트에 사용하였다. 무음구간 포함하여 이벤트별로 나눈 이유는 오디오의 연속성을 보장하면서도 독립적인 이벤트를 통해 성능을 평가하기 위해서이다.

5.3 로봇에서의 시간 소요

본 논문에서 제안된 시스템의 시간 소요를 테스트하기 위해 음향 인식 모델을 keras로 학습시키고 TensorRT[11]에서 사용할 수 있는 ONNX 모델로 변환하였다. 변환된 ONNX 모델의 추론은 C++ 코드로 개발되어 NVIDIA Xavier[10]에서 Tegra GPU 및 디바이스에 연결시킨 외부 마이크를 사용하여 테스트하였다.

정확한 측정을 위해 10번의 테스트 버퍼가 지나간 후 제안된 알고리즘으로 1,000번의 추론을 수행하였고 각 추론은 0.02 s에 해당하는 프레임을 96개(1.92 s) 수집하도록 하는 오디오 버퍼 핸들을 이용하여 추출된 feature를 사용하여 수행되었다. 한 프레임 feature 추출부터 한 번의 추론까지 평균적으로 6.575 ms가 소요되었다. 추론에 걸린 최대 시간은 10.646 ms으로 이를 식 (2)에 적용하여 최소 추론 주기 s는 1로 얻을 수 있다. 즉, 제안된 시스템은 최대 10.646 ms 이내의 시간에서 음향 이벤트를 인식할 수 있다. 추가적으로, 동시에 사용되는 다른 프로그램이 있으면 최대 추론 시간은 더 늘어날 수 있다. RF 결정 로직도 동일한 환경에서 테스트한 결과 약 0.2 ms의 시간이 소요되어 제안된 시스템의 최대 소요 시간은 10.846 ms이다. 또한 섹션 3.1에서 설명한 것처럼 추론 주기 s를 경험적으로 6으로 사용하였다.

5.4 Stream 데이터를 이용한 실험 결과

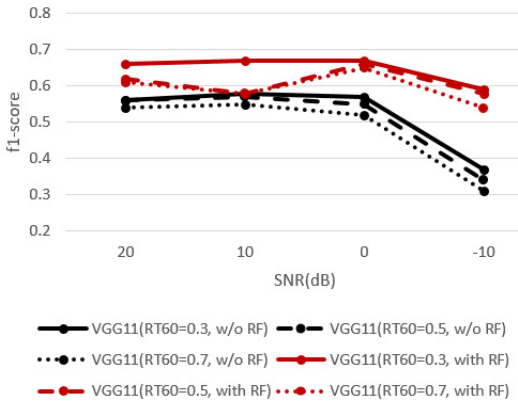
본 논문에서 사용된 Stream 데이터는 모든 이벤트 사이에 2.56 s초의 무음구간이 존재하며, 제일 처음 부분에 10초의 무음구간도 있다. 그렇게 생성된 stream 음원은 48 kHz sample rate, 96,000개 샘플 및 [5×8×2.5]의 공간 크기를 가진 rir-generator[17]로 생성된 0.3, 0.5 및 0.7의 T_{60} 특성을 가진 3개의 임펄스 응답에 필터링 된다. 필터링된 음원들은 브라운 노이즈가 SNR -10, 0, 10 및 20 dB의 4가지 경우로 더해진다. 생성된 음원은 버퍼 핸들 방식을 통하여 제안된 음향 인식 시스템에 전송된다. 음향인식 시스템에서 feature가 추출되고 모든 추론 주기 s마다 모델에서 음향 인식 확률 벡터를 추론한다.

여러 환경으로 믹싱된 stream 음원들을 사용한 실험 결과를 통해 각 추론에 대한 f1-score를 계산하였다. 정확한 비교 결과를 얻기 위해 복잡한 음향 환경의 특성을 제한하여 생성된 stream 음원들을 사용하였다. [Table 3]은 결정 로직이 있는 경우와 없는 경우를 비교한 실험 결과를 보여준다. 결정 로직을 사용하지 않는 경우에는 다른 일반적인 음향 인식 시스템들과 마찬가지로 가장 높은 확률 값을 갖는 클래스를 인식된 이벤트로 결정하였다.

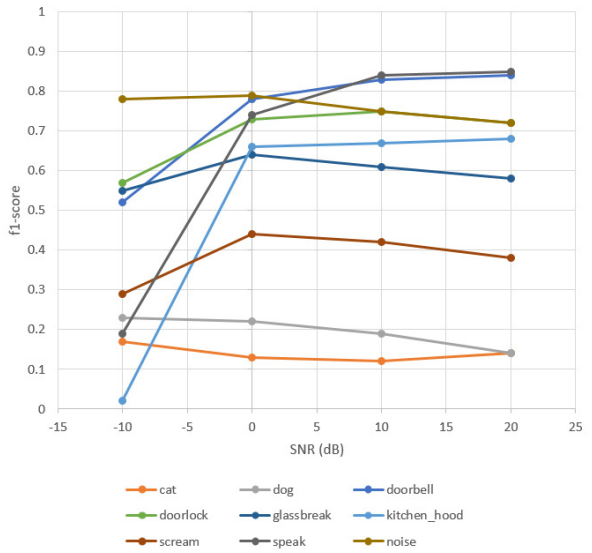
[Fig. 3]에서는 T_{60}이나 잡음의 세기에 따른 모델의 성능을 보여주고 있다. 잡음의 파워가 커질수록 성능이 떨어지는 경

[Table 3] F1-scores of experiments in various environments

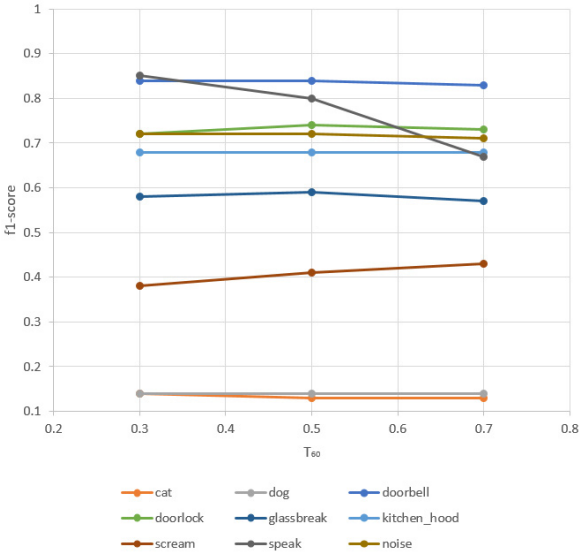
SNR (dB)	20	10	0	-10
Without RF (T_{60}=0.3)	0.56	0.58	0.57	0.37
Without RF (T_{60}=0.5)	0.56	0.57	0.55	0.34
Without RF (T_{60}=0.7)	0.54	0.55	0.52	0.31
With RF (T_{60}=0.3)	0.66	0.67	0.67	0.59
With RF (T_{60}=0.3)	0.62	0.58	0.66	0.58
With RF (T_{60}=0.3)	0.61	0.58	0.65	0.54



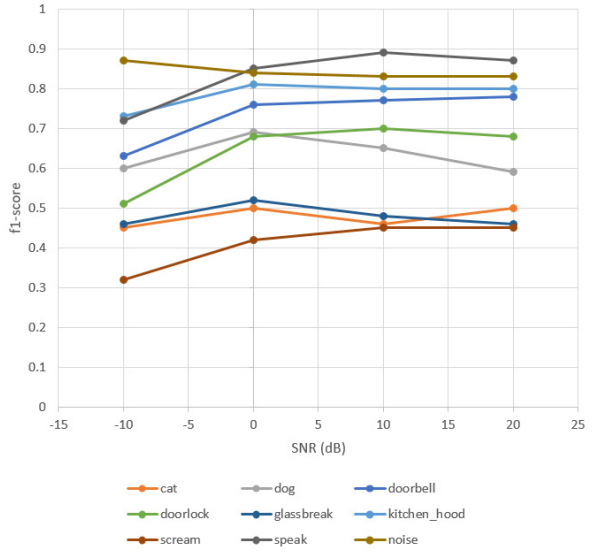
[Fig. 3] Experimental results with or without random forest



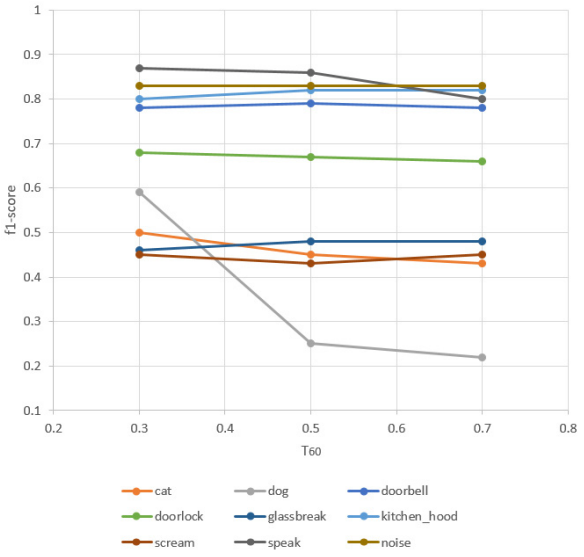
[Fig. 4] Experimental results of each classes without random forest decision logic in 0.3 T_{60}



[Fig. 6] Experimental results of each classes without random forest decision logic in 20 dB SNR



[Fig. 5] Experimental results of each classes with random forest decision logic in 0.3 T_{60}



[Fig. 7] Experimental results of each classes with random forest decision logic in 20 dB SNR

향이 있지만 10 dB, 20 dB SNR 환경에서는 DB 증강의 영향으로 인해 해당 경향성이 거의 없음을 볼 수 있다. 또한, 제안된 결정 로직이 전반적으로 모델의 성능을 약 20% 향상시키는 것을 볼 수 있다.

환경 변화에 따른 성능 변화를 클래스별로 관찰하기 위해 [Fig. 4], [Fig. 5]에서는 0.3 T_{60}을 사용하는 여러 SNR 환경에서 결정 로직이 없을 때와 RF 결정 로직을 사용할 때의 성능 변화를 보여준다. 두 그래프를 비교했을 때 'kitchen hood', 'scream', 'speak'등 긴 소리가 있는 일부 클래스 외에도'cat', 'dog'와 같은 짧은 소리 클래스의 f1-score도 증가했다. [Fig. 6], [Fig. 7]에서는 SNR을 20 dB로 고정하였을 때, 다양한 T_{60} 환경에서 제안된 결정 로직에 따른 성능 변화를 보여준다. [Fig. 4], [Fig. 5]에서와 마찬가지로 'cat', 'dog' 클래스 성능이 약 30~40% 증가하였다. [Figs. 4~7]에서 제안된 시스템은 T_{60}과 SNR이 악화되더라도 여러 클래스들의 성능이 개선되는 것을 보여준다.

6. 결 론

본 논문에서는 로봇을 위한 실시간 음향 인식 시스템을 RF 결정 로직을 이용하여 제안하였다. 제안된 시스템은 멜스케일

프레임 정규화를 사용하여 실시간 오디오 버퍼를 처리하여 음향 인식 모델에 feature로 전달한다. 음향 인식 모델은 딥러닝 추론을 통해 음향 이벤트 확률 벡터를 생성하고 생성된 벡터는 RF 결정 로직에 입력으로 사용된다. RF 알고리즘은 로봇 어플리케이션에 사용되는 이벤트 안에서 인식된 이벤트를 결정하게 된다. 제안된 시스템은 다양한 SNR과 T_{60}에 따라 로봇의 환경이 변하더라도 좋은 성능을 보였다. 때문에, 제안된 실시간 음향 인식 시스템은 음향 환경이 매우 다양하게 변하는 로봇의 어플리케이션에서 유용하게 사용될 수 있다.

References

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Vision and Pattern Recognition*, 2015, DOI: 10.48550/arXiv.1409.1556.

[2] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, 2017, DOI: 10.1109/ICASSP.2017.7952132.

[3] S. Suh, S. Park, Y. Jeong, and T. Lee, "Designing Acoustic Scene Classification Models with CNN Variants," *DCASE 2020 Challenge*, 2020, [Online], https://dcase.community/documents/challenge2020/technical_reports/DCASE2020_Suh_101.pdf

[4] H. Seo, J. Park, and Y. Park, "Acoustic scene classification using various pre-processed features and convolutional neural networks," *DCASE 2019 Challenge*, 2019, [Online], https://dcase.community/documents/challenge2019/technical_reports/DCASE2019_Seo_72.pdf

[5] T. K. Ho, "Random decision forests," *3rd International Conference on Document Analysis and Recognition*, Montreal, QC, Canada, 1995, DOI: 10.1109/ICDAR.1995.598994.

[6] C.-Y. Yu, H. Liu, and Z.-M. Qi, "Sound Event Detection Using Deep Random Forest," *DCASE 2017 Challenge*, 2017, [Online], https://dcase.community/documents/challenge2017/technical_reports/DCASE2017_Yu_162.pdf

[7] I. McLoughlin, H. Zhang, Z. Xie, Y. Song, and W. Xiao, "Robust sound event classification using deep neural networks," *IEEE/ACM Transactions On Audio, Speech, And Language Processing*, vol. 3, no. 3, March, 2015, DOI: 10.1109/TASLP.2015.2389618.

[8] I. Ozer, Z. Ozer, and O. Findik, "Noise robust sound event classification with convolutional neural network," *Neurocomputing*, vol. 272, no. 10, pp. 505-512, Jan., 2018, DOI: 10.1016/j.neucom.2017.07.021.

[9] K. Wang, J. Zhang, S. Sun, Y. Wang, F. Xiang, and L. Xie, "Investigating generative adversarial networks based speech dereverberation for robust speech recognition," *Interspeech 2018*, 2018, DOI: 10.21437/Interspeech.2018-1780.

[10] J. Lee, D. Lee, H.-S. Choi, and K. Lee, "Room adaptive conditioning method for sound event classification in reverberant environments," *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, 2021, DOI: 10.1109/ICASSP39728.2021.9413929.

[11] NVIDIA, "Jetson AGX Xavier Developer Kit," [Online], https://developer.nvidia.com/embedded/jetson-agx-xavier-developer-kit, Accessed; May 27, 2022.

[12] NVIDIA, "TensorRT," [Online], https://developer.nvidia.com/tensorrt, Accessed: May 27, 2019.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, DOI: 10.1109/ICCV.2015.123.

[14] P. Harar, R. Bammer, A. Breger, M. Dörfler, and Z. Smekal, "Improving Machine Hearing on Limited Data Sets," *2019 11th International Congress On Ultra Modern Telecommunications And Control Systems And Workshops (ICUMT)*, Dublin, Ireland, 2019, DOI: 10.1109/ICUMT48472.2019.8970740.

[15] D. Morawiec, "sklearn-porter," [Online], https://github.com/nok/sklearn-porter, Accessed: May 27, 2022.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *3rd International Conference on Learning Representations*, 2015, DOI: 10.48550/arXiv.1412.6980.

[17] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal Acoustic Society of America*, vol. 65, no. 4, 1979, DOI: 10.1121/1.382599.

송 주 만

2010 포항공과대학교 전자전기공학과(학사)
2012 포항공과대학교 정보전자융합공학부(석사)
2017 포항공과대학교 정보전자융합공학부(박사)
2017~현재 LG전자 선임연구원

관심분야: Sound Event Detection, AI, Adaptive Filter, Sound Signal Processing, Active Noise Control

박 용 진

2007 서강대학교 전자공학과(학사)
2009 서강대학교 전자공학과(석사)
2010~현재 LG전자 책임연구원

관심분야: Sound Signal Processing, Active Noise Control, Robot Audio Front-end

김 창 민

2019 서강대학교 컴퓨터공학과(학사)
2021 서강대학교 컴퓨터공학과(석사)
2021~현재 LG전자 연구원

관심분야: Sound Event Detection and Classification, Semi-Supervised Learning, Speech Recognition

이 서 영

2017 서강대학교 전자공학과(학사)
2019 서강대학교 전자공학과(석사)
2019~현재 LG전자 선임 연구원

관심분야: Robot Audio Front-end, Array Microphone Signal Processing, Speech Recognition

김 민 욱

2009 서강대학교 전자공학과(학사)
2011 서강대학교 전자공학과(석사)
2016 서강대학교 전자공학과(박사)
2016~현재 LG전자 책임연구원

관심분야: Speech Enhancement, Robot Audio Front-end, Array Microphone Signal Processing

손 정 관

2017~현재 LG전자 책임 연구원

관심분야: Sound Event Detection and Classification, Audio Signal Processing, Speech Recognition