2020 연구논문

과학영재학교 대전과학고등학교 DAEJEON SCIENCE HIGH SCHOOL FOR THE GIFTED

딥러닝 기반 VAD 모델을 이용한 잡음이 섞인 음성데이터의 전처리와 DoA에 대한 연구

A study on noise-robust DoA estimation model using deep learning based VAD applicable to speech data

장 수 한
(Su-Han Jang)

Abstract : With the advent of Fourth Industrial Revolution, the use of AI is required in disaster situation. In this social context, this study establishes a CNN model that recognizes whether the voice exists according to 2-channel speech data mixed with noise at intervals of 20 degrees and accurately estimates the angle at which the voice origin is located. The accuracy of the model was enhanced by the augmentation that introduces various kinds of noise, and the performance of introducing Voice Activity Detection (VAD) to the small interval for estimating angles was compared to applying only Direction Of Arrival (DoA) for each voice file based on its full length without dividing sections. Noise Augmentation leads to a 2.5% improvement in acuracy for DoA only model. Accuracy of the VAD model was found to be about 94 to 95%, and the F1 core was 0.9 to 0.99. Accuracy was increased by 4.3% to 6.5% by combining the VAD and DoA, resulting in significant performance improvement. In particular, combining VAD and DoA in the absence of data augmentatin contributed to a significant increase in acuracy by 6.5%.

Keywords : CNN based VAD & DoA model, data augmentation, 2-channel speech data, noise robustness, mel spectrogram

Ⅰ. 서론

1. 연구 목적

본 연구는 20도 간격의 잡음이 섞인 2-channel 음성 데이터로부터 음성이 존재하는 정확한 구간을 추정하고 이를 바탕으로 음성 발원지가 위치한 각도를 정확하게 추정하는 것을 목적으로 한다. 특히 다양한 종류의 잡음을 도입하는 augmentation을 통해 CNN 모델의 정확도를 높였으며 각도 추정을 위해 구간별 VAD (Voice Activity Detection)를 도입하는 경우와 구간을 나누지 않고 각 음성파일마다 전체길이를 바탕으로 DoA (Direction Of Arrival)를 적용하는 경우에 대한 성능을 비교하였다.

2. 연구 동향 및 수준

4차 산업혁명 시대가 도래함에 따라 재난상황에서 AI의 활용도가 요구되고 있다. 특히 대형복합재난의 효과적인 대응을 위하여 재난 상황 인식, 인명구조 등에 활용가능한 AI에 대한 연구도 진행되고 있다. [1] [2] [3] 재난 상황에서 드론으로 취득한 구조요청을 인식하여 음성의 발원지를 추정하는 DoA 연구의 필요성이 제시되고 있는 한편 (인공지능 R&D 그랜드 챌린지 대회. https://www.ai-challenge.kr/), 잡음의 종류에 강건하고 낮은 Signal-to-noise ratio (SNR)의 데이터에 대한 적응력 향상이 요구되고 있고, 음성 신호가 존재할 때 그 위치를 특정화하는 연구가 진행되고 있다. [4] 이러한 맥락 속에서 본 연구는 2-channel 음성에 대해 다양한 잡음 환경에 대한 고려와 VAD model의 적용을 바탕으로 음원의 발원지를 추정하고자 한다.

3. 이론적 배경

3.1 Voice Activity Detection (VAD)

VAD는 사람의 음성의 유무를 판단하는 음성처리를 의미한다. 오디오의 비언어구간에서 특정 프로세스를 비활성화시키기 위한 목적으로 활용된다. 활용되는 방법론으로는 먼저 음성이 존재하는 경우 상대적으로 신호의 에너지가 높다는 점을 이용하여 입력 신호의 에너지가 일정 수준을 넘는 경우 음성이 존재한다고 판단하는 Energy threshold method, 노이즈의 zero crossing rate가 음성 신호에 비해 크게 나타나는 점을 이용한 Zero Crossing Rate Method 등의 전통적인 방식들이 존재한다.

3.2 Direction of Arrival (DoA)

일반적으로 DoA란 신호가 도달하는 방향을 말하며 본 연구에서 DoA는 음성 신호가 발생한 발원 방향을 의미하며 평면상의 위치만을 고려하여 하나의 각도로 표현하고 있다.

3.3 Signal-to-noise ratio (SNR)

SNR은 잡음의 power에 대한 신호의 power의 비로 $SNR = \frac{P_{Signal}}{P_{Noise}}$와 같다. SNR_{dB}는 $SNR_{dB} = 10\log_{10}\frac{P_{Signal}}{P_{Noise}}$로 정의하고 있다.

3.4 Mel Spectrogram

Mel spectrogram은 음성 데이터의 시간에 따른 신호를 주파수 성분으로 분해하여 나타낸 것으로 짧은 시간 구간에

대하여 푸리에 변환 (Short-time Fourier transform, STFT)을 거치고 mel filter bank를 적용시켜 얻을 수 있다. Mel spectrogram은 시간에 따른 각 주파수 성분의 세기를 표현한다. 그림 1. 은 각각 좌, 우 채널에 대해 사람의 구조요청 목소리를 mel spectrogram으로 표현한 것이다. 진하게 보이는 나란한 여러 개의 띠는 사람의 목소리를 표현하고 있으며 약간의 저주파 영역 바람소리가 포함되어 있다. 음원이 좌측방향에서 발생한 경우로 왼쪽 채널의 신호의 세기가 더 큰 것을 알 수 있다.

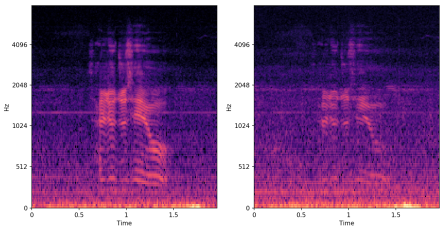

그림 1. 각각 좌, 우 채널에 대해 구조요청 소리를 mel spectrogram으로 표현한 것

3.5 F1 Score

F1 score는 이진분류 모델의 성능을 평가하는 지표 중 하나로 데이터가 불균일한 경우에도 recall과 precision의 조화평균을 통해 객관적인 평가가 가능하다. 그림 2. 와 같이 실제 정답과 분류 결과의 조합에 따라 4가지 영역으로 나눌 수 있으며 편의상 약자를 이용해 TP, FP, FN TN으로 표현하기로 하자. 이때 정밀도(precision) $= \frac{TP}{TP+FP}$, 재현율(recall) $= \frac{TP}{TP+FN}$로 정의된다. 그리고 F1 score는 이들의 조화평균인 $\frac{2\times preision\times recall}{preision+recall}$로 정의된다.

		실제 정답	
		True	False
분류 결과	True	True Positive	False Positive
	False	False Negative	True Negative

그림 2. 분류 결과와 실제 정답에 대한 TP, FP, FN, TN

II. 연구과정 및 방법

1. 데이터셋 구축

1.1 음성데이터 수집

반향음이 존재하지 않는 야외에서 0도에서 180도 까지 20도 간격으로 10가지 방향 중 한 방향으로부터 사람의 구조요청 목소리가 들려오는 상황을 가정해 동일한 길이(2초)의 음성데이터를 취득했다. 취득과정에서는 2-channel 스테레오 녹음기를 사용했으며 각 각도별로 116개씩 총 1160개의 음성데이터를 수집했다. sampling rate은 48kHz이다.

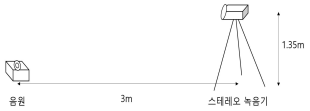

그림 3. 녹음환경

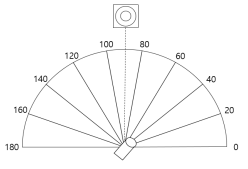

그림 4. 40도 방향 녹음 과정의 평면도

1.2 VAD 학습을 위한 data labeling

수집한 음성데이터에 대하여 사람의 음성이 포함된 구간에 대한 ground truth를 아래의 과정을 통해 제작한다.

1) High pass filter를 통해 취득한 음성 데이터에 존재하는 저주파 영역의 바람소리를 제거한다.
2) 충분히 잡음이 배제된 음성 데이터임을 가정하여 Energy threshold method를 적용한다. 취득한 데이터의 평균 energy의 1.4배 (실험적으로 결정)의 threshold를 초과하는 50ms길이의 소구간들을 labeling 한다.
3) 음성이 포함된 구간의 경계 부분을 검증한다.

1.3 잡음이 포함된 데이터 제작 (data augmentation)

위에서 수집한 음성데이터에 잡음을 추가하는 과정을 진행한다. VAD 및 DoA 모델의 효과적인 학습을 위하여 서로 다른 14가지 종류의 잡음과 5가지 수준의 잡음의 세기(noise level)를 적용하여 잡음이 포함된 데이터를 제작하였다. 동일한 세기의 잡음은 종류에 무관하게 각각의 평균 진폭이 일정하도록 정규화하였다. 이제, 1에서 수집한 각각의 음성데이터에 대하여 [예]와 같은 70가지 조합의 잡음을 추가한 음성데이터를 생각할 수 있다.

[예] (잡음 1, 세기1), (잡음1, 세기2), ... , (잡음 14, 세기4), (잡음14, 세기5)

이들 중 음성데이터에 드론소리1을 $SNR_{dB} = -9.056$ (음성과 드론소리1의 비율을 SNR_{dB}로 표현함)로 첨가한 데이터셋을 "원본 데이터"로 간주하였다. (not augmented dataset). 그리고 드론소리1의 잡음의 세기를 달리하는 것 (4종류)과 드론소리1을 대신하여 나머지 13가지 종류의 잡음을 5가지 수준의 세기로 조합하는 것을 data augmentation 방법으로 채택하였다.

수집한 14가지 종류의 잡음은 표 1.과 같고 모두 2-channel audio이며 sampling rate은 48kHz이다. 본 연구에서는 표 2.와 같이 5가지 수준의 잡음의 세기를 사용하고 있다. 표 2.는 각각의 SNR_{dB}의 값을 나타내고 있다. 원본 데이터에 해당되는 조합은 표 1., 표 2.에 '(원본)' 으로 별도

로 표시되어 있다.

표 1. 잡음의 종류

잡음의 종류	
드론소리1 (원본)	드론소리2
말벌소리	귀뚜라미소리
새소리	시냇물소리
지진소리	돌풍소리
빗소리	천둥소리
바람소리	바람, 나뭇잎 소리
차소리	경적소리

표 2. 잡음의 세기

잡음의 세기 (SNR_{dB})	
세기1	-20.988
세기2 (원본)	-9.056
세기3	-3.248
세기4	-0.024
세기5	2.139

2. 제작한 데이터 셋에 대한 모델 제안 및 학습

2.1 데이터 전처리 및 feature 추출

1) High pass filter

저주파 소음이 집중적으로 위치한 저주파 영역의 신호를 high pass filter를 이용해 제거한다.

2) Acoustic feature 추출

1)의 speech enhancement를 마친 음성 데이터의 acoustic feature로 mel spectrogram을 사용한다.

2.2 VAD 학습

2초 길이의 2-channel 음성파일을 50ms 간격으로 나누어 각각의 mel spectrogram을 학습데이터로 이용하였다. 학습에 사용된 데이터 각각의 크기는 (2×80×6) 이다. Conv2D를 기반으로 결정된 VAD 모델의 하이퍼 파라미터는 5-fold cross validation으로 결정되었으며 자세한 구조는 표 3.과 같다. 학습은 RTX 2070 super GPU 1개에서 진행했으며. 음성이 포함되지 않은 구간과 포함된 구간의 비율은 0.66777188 : 0.33222812이다. 이를 보정하기 위해 이들의 역수비로 class weight를 설정하였다. Optimizer Adam의 learning rate는 3e-4로 하였고, loss function은 weighted cross entropy loss를 사용했다. 평가지표로는 loss, accuracy와 더불어 F1 score를 사용했다.

표 3. Conv2d VAD model의 구조

Layer	Output Shape	Parameter #
Conv2d-1	[-1, 72, 80, 6]	1368
AvgPool2d-2	[-1, 72, 40, 3]	0
ReLU-3	[-1, 72, 40, 3]	0
BatchNorm2d-4	[-1, 72, 40, 3]	144
Dropout2d-5 (0.2)	[-1, 72, 40, 3]	0
Conv2d-6	[-1, 144, 40, 3]	93,456
AvgPool2d-7	[-1, 144, 20, 1]	0
ReLU-8	[-1, 144, 20, 1]	0
BatchNorm2d-9	[-1, 144, 20, 1]	288
Dropout2d-10 (0.25)	[-1, 144, 20, 1]	0
Conv1d-11	[-1, 144, 38]	31,248
ReLU-12	[-1, 144, 38]	0
BatchNorm1d-13	[-1, 144, 38]	288
Linear-14	[-1, 72]	10,440
ReLU-15	[-1, 72]	0
BatchNorm1d-16	[-1, 72]	144
Dropout-17 (0.5)	[-1, 72]	0
Linear-18	[-1, 2]	146
Total params: 137,522 (모두 학습이 가능한 parameter)		

2.3 DoA 학습

DoA 학습의 경우 2가지 방법으로 나누어 성능을 평가하였다. 첫번째 방법은 2초 길이의 각각의 2-channel 음성데이터 (2×80×244)를 바탕으로 Conv2d 기반 DoA 모델로 학습하는 것이며 두번째 방법은 2초 길이의 음성파일을 50ms 간격으로 나누어 분할된 (2×80×6)음성데이터를 바탕으로 VAD 모델이 음성이 존재하는 것으로 예측한 경우에 대하여 Conv2d 기반 DoA 모델로 학습시키며 각 음성파일에 대해 추정한 각도의 최빈값을 산출하여 첫번째 방법과 성능을 비교하는 것이다. DoA 모델의 하이퍼 파라미터는 5-fold validation으로 결정했으며 자세한 구조는 표 4.와 같다. 학습은 RTX 2070 super GPU 1개에서 진행하였으며 각 각도 클래스별로 데이터의 수는 균일하기 때문에 loss와 accuracy만을 평가지표로 사용했다. Optimizer Adam의 learning rate는 3e-4로 하였고 loss function은 cross entropy loss를 이용하였다.

표 4. Conv2d DoA model의 구조, 마지막 인덱스의 괄호 밖의 숫자는 (2×80×244) 크기의 input이 주어진 경우 괄호 안의 숫자는 (2×80×6) 크기의 input이 주어진 경우를 표현함

Layer	Output Shape	Parameter #
Conv2d-1	[-1, 128, 80, 244 (6)]	2,432
AvgPool2d-2	[-1, 128, 40, 122 (3)]	0
ReLU-3	[-1, 128, 40, 122 (3)]	0
BatchNorm2d-4	[-1, 128, 40, 122 (3)]	256
Dropout2d-5 (0.2)	[-1, 128, 40, 122 (3)]	0
Conv2d-6	[-1, 256, 40, 122 (3)]	295,168
AvgPool2d-7	[-1, 256, 20, 61 (1)]	0
ReLU-8	[-1, 256, 20, 61 (1)]	0
BatchNorm2d-9	[-1, 256, 20, 61 (1)]	512
Dropout2d-10 (0.25)	[-1, 256, 20, 61 (1)]	0
Conv1d-11	[-1, 256, 2438 (38)]	98,560
ReLU-12	[-1, 256, 2438 (38)]	0
BatchNorm1d-13	[-1, 256, 2438 (38)]	512
Linear-14	[-1, 128]	32,896
ReLU-15	[-1, 128]	0
BatchNorm1d-16	[-1, 128]	256
Dropout-17 (0.5)	[-1, 128]	0
Linear-18	[-1, 10]	1,290
Total params: 431,882 (모두 학습이 가능한 parameter)		

VAD model과 DoA model의 train set과 valid set은 서로 똑같이 분리하였으며 "드론소리1" 의 잡음이 섞인 음성데이터를 기본으로 하여 train set에만 나머지 13가지 종류의 잡음과 5가지 종류의 noise level에 대하여 data augmentation을 진행하였다.

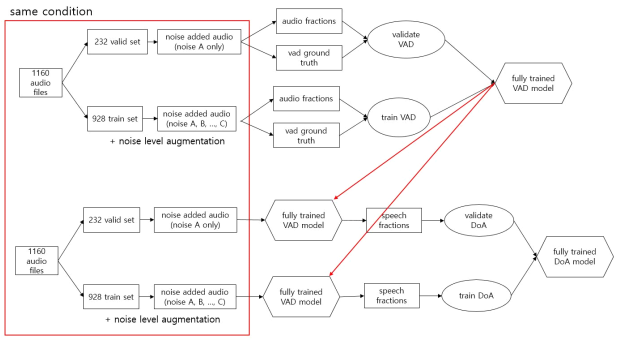

그림 5. VAD를 이용해 DoA를 학습하는 과정의 모식도

III. 연구 결과

1. Conv2d based VAD의 성능

1.1 Augmentation을 적용한 경우

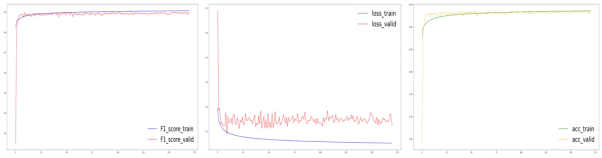

그림 6. Augmentation을 적용한 경우 VAD의 성능 왼쪽부터 F1 score, loss, accuracy

Augmentation을 적용한 경우 Conv2d 기반 VAD 모델의 accuracy는 약 93.6% 정도이며 F1 score는 0.902로 나타났다.

1.2 Augmentation을 적용하지 않은 경우

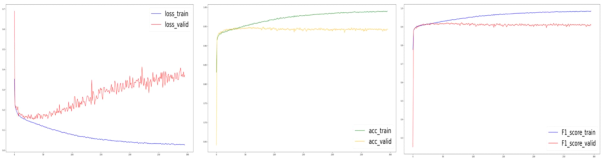

그림 7. Augmentation을 적용하지 않은 경우 VAD의 성능 왼쪽부터 F1 score, loss, accuracy

Augmentation을 적용하지 않은 경우 Conv2d 기반 VAD 모델의 accuracy는 약 94.9% 정도이며 F1 score는 0.923으로 나타났다. augmentation을 적용한 경우에 비해 성능이 높게 나온 것은 주어진 잡음의 종류와 세기가 다양해 학습하기 어려웠기 때문이라고 생각되나 성능의 차이는 accuray는 약 1%, F1 score의 경우 0.02 수준으로 두드러지게 나타나지는 않았다. 한편 augmentation을 적용한 경우와는 달리 특정 epoch 이상에서는 극심하게 과학습이 진행되는 것을 확인할 수 있었다. 이는 단일 종류의 잡음으로 학습하는 경우 과도한 overfitting이 발생할 수 있음을 시사한다.

2. DoA 단독으로 학습한 경우와 VAD, DoA를 함께 이용하여 학습한 경우 성능의 차이

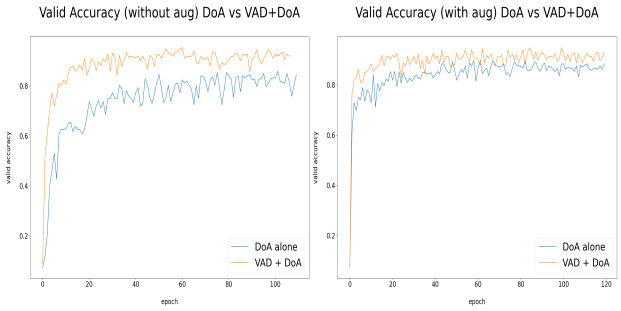

그림 8. DoA와 VAD + DoA의 valid accuracy 비교

DoA를 단독으로 사용하여 학습한 경우에 비해 VAD와 DoA를 함께 이용해 음성이 존재하는 소구간들로 학습시키는 방법의 accuracy가 유의미하게 높은 것을 확인하였다. 두 개의 accuracy를 비교하기 위해서 VAD와 DoA를 함께 사용한 경우 각 음성파일에 대하여 소구간들의 추정치의 최빈값으로 모델의 예측값을 산출하였다. 그 결과 89~91% 수준의 정확도를 95~96%의 정확도로 향상시키는데 성공하였다. 특히 잡음의 종류와 세기에 따른 data augmentation을 적용하지 않은 경우에 대해서 약 6.5% 가량의 확연한 향상이 나타났으며 data augmentation을 적용한 경우에는 4.3% 가량의 향상이 나타났다.

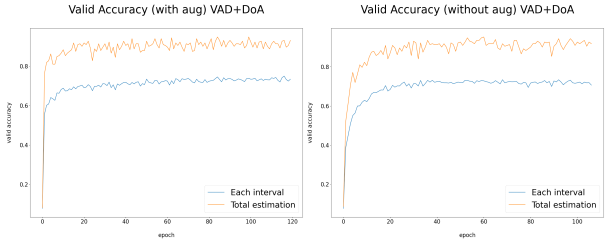

그림 9. VAD + DoA의 valid accuracy

한편 VAD와 DoA를 함께 사용한 경우에서 소구간들에 대한 accuracy는 (소구간 각각을 정확하게 맞출 확률) 은 73~75% 수준으로 낮게 측정되었는데 이는 음성구간들의 시간간격이 매우 좁아 학습이 다소 어렵고 따라서 각각에 대한 accuracy는 비교적 낮은 양상을 띠는 것이라고 생각한다.

3. Augmentation의 적용에 따른 성능의 변화

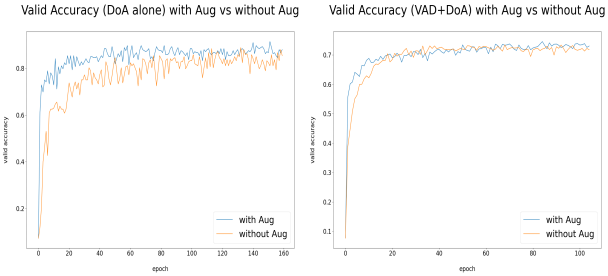

그림 10. Augmentation의 적용 여부에 따른 valid accuracy

잡음의 종류와 세기를 이용한 data Augmentation의 적용에 대한 성능의 변화를 추가적으로 측정해보았다. 그 결과 DoA를 단독으로 이용하는 모델의 경우 2.6%의 유의미한 accuracy의 향상이 관찰되었다. 하지만 VAD와 DoA를 동시에 활용하는 모델의 경우 유의미한 accuracy의 차이가 드러나지 않았다. 이는 data augmentation의 목적이 다양한 잡음 환경이 존재하는 상황에서 능동적으로 잡음이 아닌 소리에 대한 각도를 추정하는 모델의 능력을 향상시키기 위함인데 VAD를 도입하는 것만으로 관련된 성능이 충분히 보장되기 때문일 것이다.

4. Augmentation의 적용에 따른 성능의 변화

표 5. 각 DoA 추정의 실험방법별 accuracy와 loss를 나타냄. *표시가 되어 있는 경우 분할된 구간에 대한 accuracy를 나타내며 그 위의 데이터는 각 음성파일별로 추정된 각도들의 최빈값으로 결과를 산출한 경우의 accuracy를 나타냄. 괄호안의 숫자는 최대/최소에 도달한 epoch를 의미함.

DoA	Not augmented		Augmented	
	DoA alone	VAD + DoA	DoA alone	VAD + DoA
Max accuracy (train)	99.569% (162)	100% (86) *90.192% (104)	96.079% (252)	100% (3) *85.675% (361)
Max accuracy (valid)	88.793% (132)	95.259% (60) *73.426% (68)	91.379% (151)	95.69% (175) *75.297% (304)
Min loss (train)	0.05343 (162)	0.26002 (104)	0.11721 (252)	0.37541 (361)
Min loss (valid)	0.28103 (117)	0.6188 (36)	0.23052 (198)	0.58941 (325)

표 5.는 각 DoA 추정의 실험방법별 최대 accuracy와 최소 loss를 정리한 것이다. 음성데이터가 각각의 각도 클래스에 대해 균일하게 분포하기 때문에 accuracy를 주요 평가지표로 사용하였다. DoA를 단독으로 사용한 경우 88.793%

의 정확도를 보였으며 잡음의 종류와 세기에 따라 heavy augmentation을 한 경우 정확도는 약 2.5%가량 증가하여 91.379%를 기록했다. VAD를 거친 후 음성이 존재하는 구간에 대해 DoA를 적용하는 방식으로 학습한 결과 분할된 음성 각각에 대한 정확도는 73~75%의 비교적 낮은 결과를 보였으나 이들을 종합하여 각 음성파일에 대해 추정치의 최빈값으로 최종적인 DoA를 추산한 결과 augmentation이 없는 조건에서는 95.26%, 있는 조건에서는 95.69%로 나타났다. 요약하면 VAD와 DoA를 종합하는 방식으로 정확도를 4.3%~6.5%가량 높일 수 있었고 유의미한 성능향상을 이끌어냈다고 볼 수 있다.

표 6. 각 VAD 추정의 실험방법별 F1 score, accuracy와 loss를 나타냄. 괄호안의 숫자는 최대/최소에 도달한 epoch를 의미함.

VAD	Not Augmented	Augmented
Max F1 score (train)	0.98584 (299)	0.90782 (172)
Max F1 score (valid)	0.92269 (81)	0.90162 (52)
Max accuracy (train)	99.052% (299)	93.649% (172)
Max accuracy (valid)	94.883% (81)	93.623% (52)
Min loss (train)	0.02461 (295)	0.15199 (168)
Min loss (valid)	0.15257 (33)	0.18801 (9)

표 6.은 augmentation 유무에 따른 VAD 모델의 최대 F1 score, accuracy 와 최소 loss를 표현한 것이다. 음성이 존재하는 길이와 존재하지 않는 길이의 비율은 0.332 : 0.668로 음성의 유무에 대하여 데이터가 불균일하기 때문에 F1 score를 주요 평가지표로 사용하였다. VAD 모델은 약 94~95%의 정확도를 보이고 있으며 F1 score는 0.9~0.99 수준으로 나타났다.

Ⅳ. 결론

본 연구에서는 2-channel 음성데이터에 잡음의 종류와 세기를 통해 data augmentation을 진행하였음며 Conv2d 기반 VAD 모델을 이용하여 음성이 존재하는 구간을 먼저 인식하고 이를 바탕으로 Conv2d 기반 DoA 모델을 학습시켜 DoA만으로 학습시킬 때보다 성능을 향상시켰다. 그 결과 VAD 모델의 정확도는 약 94~95%, F1 score는 0.9~0.99 수준으로 나타났고 VAD와 DoA를 종합하는 방식으로 정확도를 4.3%~6.5%가량 높혀 유의미한 성능향상을 이끌어내는데 성공하였다. DoA를 단독으로 사용하는 경우 augmentation을 통해 2.5%의 accuracy 향상을 유도할 수 있었다. 특히 augmentation이 없는 상황에서 VAD와 DoA를 동시에 사용하는 것은 6.5% 가량 accuracy를 크게 높이는데 기여하였다.

Ⅵ. 참고문헌

[1] R. Lagerstorm et al., "Image Classification to Support Emergency Situation Awareness," *Frontiers in Robotics and AI*, vol. 3, no. 54, pp. 1-3, September, 2016.

[2] S. Pouyanfa et al., "Multimodal deep learning based on multiple correspondence analysis for disaster management," *World Wide Web*, pp. 1-5, September, 2018.

[3] F. Niroui et al., "Deep Reinforcement Learning Robot for Search and Rescue Applications: Exploration in Unknown Cluttered Environments," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1-2, January, 2019.

[4] S. Adavanne, A. Politis, T. Virtanen, "A Multi-Room Reverberant Dataset For Sound Event Localization And Detection" *Audio Research Group, Tampere University, Finland*, pp. 1-5, May, 2019.