실시간 음성 모니터링을 위한 오토인코더 기반 FTAE 설계 및 구현

양진환[1], 최혁순[1], 박정현[1], 김성식[2], 문남미[3]
[1]호서대학교 컴퓨터공학부 석사과정
[2]호서대학교 컴퓨터공학부 학부생
[3]호서대학교 컴퓨터공학부 교수
yjh970706@naver.com, hyuksoon2001@gmail.com, Jh.park970609@gmail.com,
sungsik001004@gmail.com, nammee.moon@gmail.com

The Design and Implementation of Autoencoder-Based FTAE for Real-Time Audio Monitoring

Jin-Hwan Yang[1], Hyuk-Soon Choi[1], Jeong-hyeon park[1],
Sung-Sik Kim[1], Nammee Moon[1]
[1]Dept. of Computer Engineering, Hoseo University

요 약

본 연구에서는 음성 전처리 기법인 푸리에 변환의 높은 시간 복잡도로 인해 많은 계산 자원을 요구한다는 단점을 보완하기 위한 FTAE(Fourier Transform Auto Encoder)를 설계하고 구현한다. FTAE는 음성 데이터를 입력으로 받아 Early Fusion 특징맵을 출력하도록 설계된 오토인코더 기반 신경망이다. 학습 결과 FTAE의 최종 Training Loss는 0.1479를 나타냈다. 기존 푸리에 변환 기반 Early Fusion 방법과의 성능 비교 실험 결과 FTAE 방법은 Accuracy 0.905, F1-Score 0.905, 탐지 소요 시간 17초의 성능을 보였다. FTAE 방법은 Early Fusion 방법에 비해 Accuracy와 F1-Score는 0.065 하락했지만, 탐지 소요 시간은 약 72배 빠른 결과를 보여주었다.

1. 서론

푸리에 변환은 음성 데이터 분석에서 널리 쓰이는 전처리 기법이다[1-4]. 푸리에 변환을 기반으로 하는 Early Fusion 방법은 음성 분류에서 높은 성능을 보였다[5]. 하지만 푸리에 변환의 시간 복잡도는 $O(n^2)$, 이를 개선한 고속 푸리에 변환은 $O(n \log n)$ 이기 때문에 실시간 음성 모니터링에 적용하기엔 너무 많은 계산 자원을 요구한다[6,7].

따라서 본 연구에서는 푸리에 변환의 단점을 보완하기 위해 FTAE를 제안한다. FTAE는 푸리에 변환 기법을 오토인코더에 학습시켜 계산 자원을 절약하고 실시간 음성 모니터링을 위한 초석이 된다.

2. 관련 연구

2.1 Fourier Transform

푸리에 변환은 시간이나 공간에 대한 함수를 주파수 성분으로 분해하는 수학적 연산이다. 이는 복잡한 함수나 신호를 더 간단한 주파수 성분으로 나누어 분석하기 위해 사용된다[6,7]. 푸리에 변환은 시간 도메인의 함수를 주파수 도메인의 함수로 변환하기 때문에 시간에 따라 변화하는 신호가 어떤 주파수를 포함하는지 알 수 있다[8].

2.2 Early Fusion

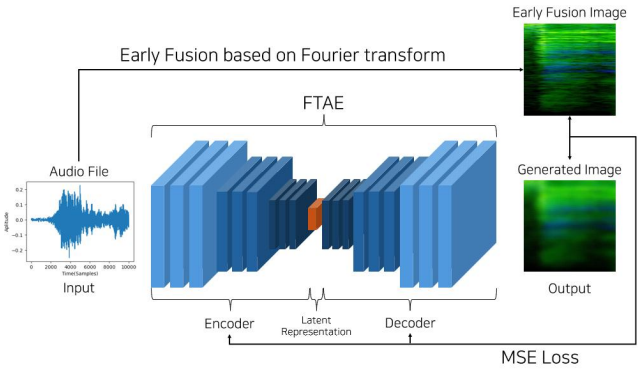
Early Fusion 방법은 하나의 음성 데이터에 세 종류의 푸리에 변환 기반 음성 특징 추출 방법(STFT, Spectrogram, Mel-Spectrogram)을 적용하고 각각 이미지의 RGB 채널에 대응하여 하나의 특징맵을 제작하는 음성 데이터 전처리 방법이다. Early Fusion 방법은 음성 분류 모델에서 하나의 특징 추출 방법만 사용했을 때보다 높은 성능을 보였다[5].

2.3 오토인코더

오토인코더는 신경망을 활용하여 입력 데이터의 효율적인 표현을 학습하는 비지도 학습 알고리즘이며 주로 데이터의 압축, 노이즈 제거, 차원 축소 등에 사용된다[9]. 오토인코더는 크게 인코더와 디코더로 구성된다. 인코더는 입력 데이터를 받아 잠재 표현으로 변환하는 역할을 한다. 이 과정에서 데이터의 중요한 특성이 추출되고 상대적으로 중요하지 않은 부분은 버려진다. 디코더는 인코더에서 생성된 잠재 공간을 입력받아 원본 데이터와 같은 차원의 출력을 생성한다. 그 후 출력과 입력의 손실을 계산하여 오토인코더를 학습한다.

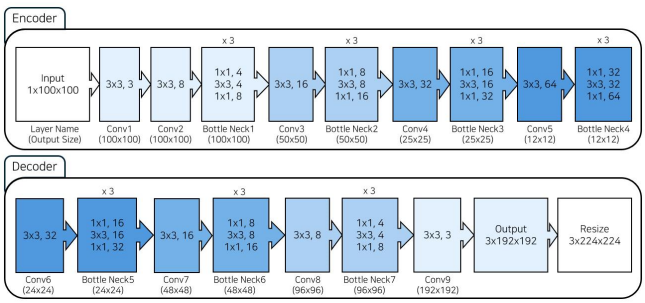
3. FTAE

FTAE는 음성 데이터를 입력으로 받아 Early Fusion 특징맵을 출력하도록 설계된 오토인코더 기반 신경망이다. 어떠한 전처리도 진행하지 않은 음성 데이터를 모델에 입력하고 인코더와 디코더를 거쳐 (3, 224, 224) 크기의 이미지를 출력한다. 그 후 출력된 이미지에 해당 음성 데이터의 Early Fusion 특징맵을 Label로 MSE Loss를 계산하여 인코더와 디코더의 역전파를 진행한다. 이 과정을 통해 오토인코더는 음성 데이터에서 Early Fusion 특징맵으로의 변환과정을 학습하게 된다. FTAE 전체 학습 과정의 개념도는 (그림 1)과 같다.



(그림 1) FTAE 학습 개념도

FTAE의 인코더는 Sample Rate가 20,050인 음성 데이터를 10,000개 신호 길이(약 0.5초)로 자르고 (100, 100)의 2차원 텐서로 변환한 뒤 입력받는다. 입력된 텐서는 총 5개의 2D Convolution Layer와 4개의 Bottle Neck Layer를 거쳐 잠재 표현으로 변환된다. 그 후 잠재 표현은 디코더를 통해 4개의 2D Convolution Layer와 3개의 Bottle Neck Layer를 거쳐 (3, 192, 192)의 이미지를 생성한다. 생성된 이미지는 Early Fusion 이미지와 MSE Loss를 계산하기 위해 (3, 224, 224)로 Resize를 진행한다. FTAE의 총 파라미터 수는 101,609개이며 구조는 (그림 2)와 같다.



(그림 2) FTAE 구조도

4. 실험

4.1 데이터셋

본 연구는 AI-Hub의 '도시 소리 데이터'와 '극한 소음 환경 소리 데이터'를 활용하여 FTAE를 학습한다. 해당 데이터셋은 교통소음(지상운송수단, 철로운송수단, 항공운송수단, 수상운송수단), 생활소음(충격, 가전, 동물, 도구), 사업장소음(공사장, 공장)등의 다양한 음성 데이터를 포함하고 있다. 이를 총 232,545개의 음성 클립으로 변환하여 활용한다. 총 데이터 규모는 <표 1>과 같다.

<표 1> FTAE 학습 데이터 분포표(단위: 개)

교통수단	공사장	공장	시설류	생활소음
68,536	45,853	46,385	27,770	44,001

또한 FTAE의 성능을 평가하기 위해 Kaggle의 'Audio Cats and Dogs'를 활용하여 음성 탐지를 진행한다. 해당 데이터셋은 반려견과 반려묘의 울음소리로 구성되어 있다. 이 중 반려견 데이터만 활용하여 실험을 진행한다. 실제 소음 환경에서 반려견의 울음소리를 탐지하는 상황을 가정하기 위해 FTAE 학습에 참여하지 않은 소음 데이터를 함께 사용한다. 총 데이터 규모는 <표 2>와 같다.

<표 2> FTAE 성능 평가 데이터 분포표(단위: 개)

반려견 울음소리	소음
2,189	1,999

4.2 실험 세부사항

FTAE는 Adam Optimizer를 활용해 0.001의 Learning Rate로 총 200 epoch 학습을 진행하며 Scheduler를 활용해 50 epoch마다 Learning Rate를 1/10로 조정한다.

그 후 FTAE의 성능 평가를 위해 Early Fusion 방법과의 비교 실험을 진행한다. 전체 4,188개의 데이터를 6:2:2 비율로 Train(2,714), Validation(736), Test(738) 데이터셋으로 나누고 약 6분 분량의 음성 데이터인 Test 데이터셋에 대한 Accuracy, F1-Score, 탐지 소요 시간을 비교한다. 탐지 모델은 사전 훈련되지 않은 ResNet50을 백본으로 사용한다. Loss Fuction으로 CrossEntropy를 사용하고 Adam Optimizer를 활용해 0.001의 Learning Rate로 총 100 epoch 학습을 진행하며 Scheduler를 활용해 20 epoch마다 Learning Rate를 1/10로 조정한다.

4.3 실험 환경

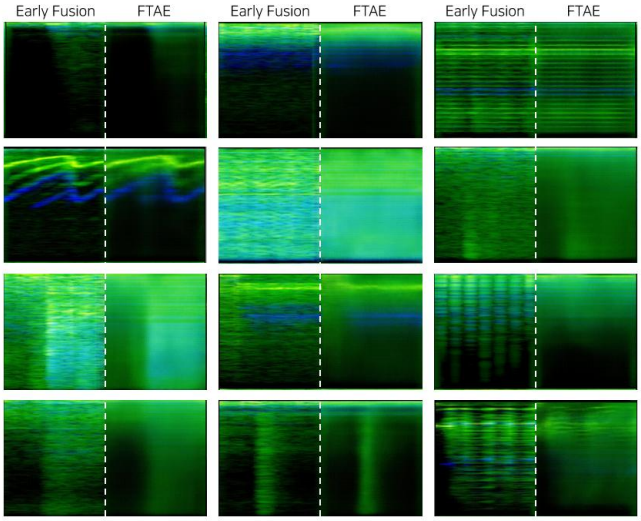
실험은 모두 같은 환경에서 진행한다. 실험 환경은 <표 3>과 같다.

<표 3> 실험 환경 사양표

유형	내용
CPU	Intel i7-11700
GPU	NVIDIA GeForce RTX 3090
Ram	64GB
CUDA	11.2
Python	3.8
PyTorch	1.7.1
Torchvision	0.8.2

4.4 실험 결과

FTAE의 최종 Training Loss는 0.1479로 나타났다. Early Fusion 특징맵과. FTAE의 최종 생성 이미지 비교는 (그림 3)과 같다.



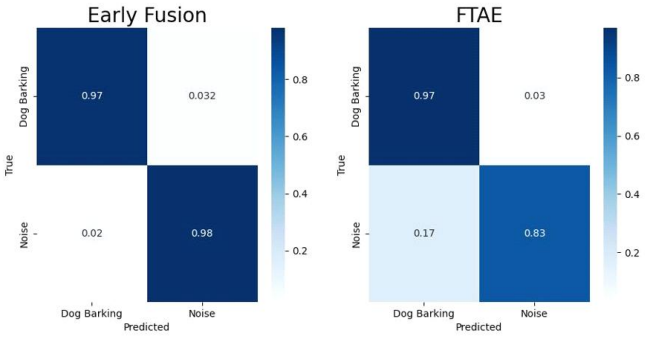
(그림 3) Early Fusion(좌) FTAE 생성 이미지(우)

(그림 3)을 보면 FTAE 생성 이미지가 Early Fusion 특징맵의 특징을 잘 묘사하고 있는 것을 확인할 수 있다. 이렇게 구현한 FTAE와 Early Fusion 방법을 활용한 반려견 울음소리 탐지 실험 결과는 <표 4>와 같다.

<표 4> 실험 결과표

Method	Accuracy	F1-Score	Time(s)
Early Fusion	0.97	0.97	1,238
FTAE	0.905	0.905	17

실험 결과 Early Fusion 방법은 Accuracy 0.97, F1-Score 0.97의 성능을 보였지만 6분 분량의 음성 데이터를 탐지하는 데 1,238초가 소요되었다. FTAE를 사용한 방법은 Accuracy 0.905, F1-Score 0.905로 Early Fusion 방법보다 0.065의 성능 하락을 보였지만 마찬가지로 6분 분량의 음성 데이터를 탐지하는 데 17초가 소요되었다. FTAE 방법이 Early Fusion 방법보다 약 72배의 탐지 속도 향상을 보였다. 각 방법을 활용한 탐지 모델의 Confusion Matrix는 (그림 4)와 같다.



(그림 4) Confusion Matrix

5. 결론

본 연구는 실시간 음성 모니터링을 위해 푸리에 변환의 단점인 높은 시간 복잡도를 개선하기 위한 FTAE를 설계하고 구현하였다. 그 후 반려견 울음소리 데이터와 소음 데이터를 활용해 FTAE와 Early Fusion의 성능 비교를 위한 실험을 진행하였다. 실험 결과 FTAE의 최종 Training Loss는 0.1479로 나타났다. 성능 비교를 위한 반려견 울음소리 탐지 모델 실험은 기존 Early Fusion 방법이 Accuracy 0.97, F1-Score 0.97, 탐지 소요 시간 1,238초를 보였고 FTAE 방법은 Accuracy 0.905, F1-Score 0.905, 탐지 소요 시간 17초를 보였다. FTAE 방법은 Early Fusion 방법에 비해 Accuracy와 F1-Score는 0.065 하락했지만 탐지 소요 시간은 약 72배 빠른 결과를 보여주었다.

6. 향후계획

본 연구에서 제안한 FTAE를 활용하여 빠른 탐지 소요 시간을 유지하며 정확도를 개선하기 위해 Early Fusion 특징맵에 특성을 고려한 Loss Filter에 관한 연구를 진행할 계획이다.

ACKNOWLEDGEMENT

본 연구는 과학기술정부통신부와 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음 (No. 2019-0-01834).

참고문헌

[1] SALAH, Euschi, et al. A Fourier transform based audio watermarking algorithm. Applied Acoustics, 2021, 172: 107652.

[2] HASSAN, KM NAIMUL; HAQUE, Mohammad Ariful. ASFNet: Audio Spectrogram Fourier Network for Efficient Medical Sound Event Detection. Authorea Preprints, 2023.

[3] BARTUSIAK, Emily R.; DELP, Edward J. Frequency domain-based detection of generated audio. arXiv preprint arXiv:2205.01806, 2022.

[4] 박정현, 고준혁, 김시웅, & 문남미. (2023). 음성 데이터 증강을 통한 3D 특징 벡터 기반 신생아 울음소리 분류. 한국컴퓨터정보학회논문지, 28(9), 47-54.

[5] 양진환, 김성식, 최혁순, 문남미, Early Fusion을 적용한 위급상황 음향 분류, 한국정보처리학회 ACK 2023, 부경대학교 대연캠퍼스, 2023, 1213-1214

[6] PAWAR, Sameer; RAMCHANDRAN, Kannan. Computing a k-sparse n-length discrete Fourier transform using at most 4k samples and O (k log k) complexity. In: 2013 IEEE International Symposium on Information Theory. IEEE, 2013. p. 464-468.

[7] DRISCOLL, James R.; HEALY, Dennis M. Computing Fourier transforms and convolutions on the 2-sphere. Advances in applied mathematics, 1994, 15.2: 202-250.

[8] SMITH, Julius O. Mathematics of the discrete Fourier transform (DFT): with audio applications. Julius Smith, 2008.

[9] BANK, Dor; KOENIGSTEIN, Noam; GIRYES, Raja. Autoencoders. Machine learning for data science handbook: data mining and knowledge discovery handbook, 2023, 353-374.