특집논문 (Special Paper)
방송공학회논문지 제30권 제3호, 2025년 5월 (JBE Vol.30, No.3, May 2025)
https://doi.org/10.5909/JBE.2025.30.3.289
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

FCM을 이용한 기계를 위한 오디오 부호화 기술 연구

변 수 빈[a)], 서 정 일[a)‡]

A Study on Audio Encoding Technology for Machines Using FCM

Subin Byun[a)] and Jeongil Seo[a)‡]

요 약

최근 딥러닝 기술의 급속한 발전은 음향 분석 등 다양한 산업 분야에서 뛰어난 성능을 보이고 있다. 그러나 기존 인간 청취 위주의 오디오 코덱은 기계가 필요로 하는 feature를 보존하기 어렵다는 한계가 존재한다. 이에 따라 본 논문에서는 FCM을 활용한 ACoM을 제안한다. 제안 방법은 오디오 신호를 스펙트로그램으로 변환한 뒤, ResNet101로 추출한 feature를 압축하여 낮은 비트레이트에서도 이상 음향을 효과적으로 감지할 수 있도록 한다. 이를 통해 불필요한 정보를 최소화하면서도 이상 음향 탐지에 중요한 핵심 feature를 효율적으로 보존하도록 설계하였다. 실험 결과, ACoM은 기존 AAC 대비 안정적이며 우수한 성능을 보였으며, 향후 대규모 데이터셋 및 딥러닝 최적화를 통해 산업 현장에서 음향 모니터링 및 이상 상태 감지에 효과적인 차세대 오디오 부호화 솔루션으로 발전할 것으로 기대된다.

Abstract

Recent advances in deep learning technology have demonstrated exceptional performance in various industrial domains, including acoustic analysis. However, traditional audio codecs designed for human hearing present limitations in preserving the features necessary for machine-based processing. In response, this paper proposes an approach called ACoM, which leverages FCM. The proposed method transforms audio signals into spectrogram and then compresses the features extracted by ResNet101 to enable effective anomaly detection at low bitrates. This design minimizes unnecessary information while efficiently retaining key features critical for anomaly detection. Experimental results indicate that ACoM achieves more stable and superior performance compared to conventional AAC. Furthermore, with the use of large-scale datasets and optimized deep learning techniques, ACoM is expected to evolve into a next-generation audio encoding solution for industrial acoustic monitoring and anomaly detection.

Keyword : ACoM, FCM, Deep Learning, ASD

a) 동아대학교 컴퓨터공학과(Dept of Computer Engineering, Dong-A University)
‡ Corresponding Author : 서정일(Jeongil Seo)
E-mail: jeongilseo@dau.ac.kr
Tel: +82-51-200-7796
ORCID: https://orcid.org/0000-0001-5131-0939
※본 연구는 교육부의 재원으로 한국연구재단의 기초연구사업(RS-2024-00394288)과 과학기술정보통신부의 재원으로 한국연구재단의 생애첫연구사업(No.RS-2023-00273349)의 지원을 받아 수행된 결과임.
· Manuscript March 26, 2025; Revised May 7, 2025; Accepted May 8, 2025.

Ⅰ. 서 론

Machine hearing 기술은 딥러닝을 활용하여 인간의 귀처럼 소리를 인식하고 분석하는 기술이다. 최근 딥러닝 발전으로, 이 기술은 공장 및 산업 현장에서 음성 인식, 소리 이벤트 감지, 오디오 분류 등 다양한 작업에 활용되고 있다[1]. 현장의 센서로 수집된 데이터를 중앙 서버나 엣지 디바이스로 실시간 전송하여 처리해야 하는 환경에서는 네트워크 대역폭과 저장 자원의 한계로 인해 데이터 압축 전송이 필수적이다. 기존 오디오 코덱은 인간 청취 중심으로 데이터를 압축하기 때문에 기계가 인식하고 처리해야 할 정보를 충분히 보존하지 못할 가능성이 있다.

이를 해결하기 위해 MPEG(Moving Picture Experts Group)은 기계 중심의 데이터 분석 성능은 유지하면서도 데이터 전송 효율을 높일 수 있는 새로운 비디오 및 오디오 압축 기술, 즉 VCM(Video Coding for Machines)[2], FCM(Feature Coding for Machines)[3,4,5], ACoM(Audio Coding for Machines)[6] 등을 개발하고 있다. 특히 ACoM은 기계가 필요로 하는 오디오의 핵심 정보만을 추출하여 압축함으로써 데이터 전송 및 저장 공간을 절약하고, 대량의 데이터를 효과적으로 처리할 수 있도록 돕는다.

본 연구에서는 1차원 오디오 신호를 2차원 스펙트로그램으로 변환한 후, FCM을 이용하여 기계가 필요로 하는 feature를 추출하여 압축하여 전송하는 접근법을 제안한다.

Ⅱ. 관련 연구

1. FCM

FCM은 영상신호를 딥러닝 네트워크로 처리하는 과정 중 생성되는 feature map을 압축 전송함으로써 딥러닝 엔진의 성능은 유지하면서 영상 데이터 전송 효율을 극대화하는 기술이다. 기존 영상 압축 방식과 달리, FCM은 딥러닝 엔진이 필요로 하는 feature 정보만을 선택적으로 인코딩함으로써 네트워크 대역폭과 전송 비용을 크게 절감한다.

FCTM(Feature Compression Test Model)은 FCM 표준화 과정 중 성능 평가 및 검증을 위해 활용되는 참조 소프트웨어(Reference Software)이다[7]. 전체 구조는 그림 1에 제시되어 있으며, FCTM은 NN-Part1, FCM Encoder/Decoder, NN-Part2로 구성된다. 먼저, NN-Part1은 입력 이미지를 받아 중간 feature map을 추출하는 역할을 수행한다. 이후 해당 feature는 FCM Encoder를 통해 Feature Reduction, Feature Conversion, 그리고 VVC 기반 내부 코

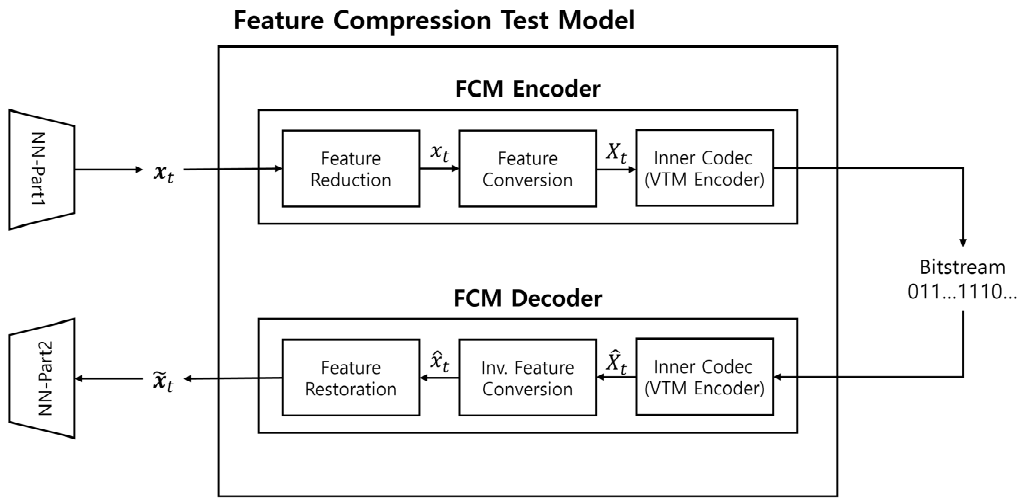

그림 1. FCM 참조 소프트웨어인 FCTM 블록도
Fig. 1. Block diagram of FCTM, the reference software for FCM

덱을 거쳐 압축된 비트스트림으로 변환된다.

압축된 비트스트림은 FCM Decoder에서 복원 과정을 거치며, 이는 feature 역변환(Inverse Feature Conversion), 복원(Feature Restoration) 단계를 포함한다. 이렇게 복원된 feature map은 NN-Part2로 전달되어, 이후 detection, segmentation, classification 등 다양한 AI 기반 task에 사용된다.

1.1 FCM Encoder

FCM Encoder는 Feature Reduction, Feature Conversion, Inner Codec 순으로 인코딩을 수행한다. Feature Reduction 단계에서는 feature 자체의 크기를 줄이는 것을 목표로 feature padding, multi-scale feature fusion, 그리고 encoding 과정을 진행한다. 이 단계에서 multi-scale feature map을 single-scale feature map으로 변환하기 위해, 먼저 feature padding을 통해 각 층에 있는 feature map을 미리 정의된 크기로 조정한다. 이후 그림 2의 FENet(Feature Fusion and Encoding Network)을 사용하여 여러 스케일의 feature map을 하나로 합치는 feature fusion과 feature map의 크기를 줄이는 encoding을 수행한다. FENet의 구조를 자세히 보면, 가장 높은 해상도를 가진 패딩된 feature인 x_{pad}^{1}은 첫 번째 인코딩 블록에 입력되어 두 번째로 높은 해상도를 가진 패딩된 feature인 x_{pad}^{2}와 동일한 해상도를 가진 latent feature인 y^{1}을 생성한다. 이후 y^{1}와 x_{pad}^{2}는 후속 인코딩 블록을 통해 연결되고 처리된다. 이를 통해 최종적으로 320 차원을 가지는 single-scale feature map이 생성되며, 마지막으로 gain unit을 거쳐 최적화된 latent 표현으로 변환된다.

Feature Conversion 단계에서는 packing, quantization range derivation, normalization, quantization 순으로 feature를 변환하고 압축한다. 먼저, packing 과정에서는 3차원 텐서로 구성된 feature y를 하나의 feature 프레임으로 변환하여 y_p로 패킹한다. 이후 quantization range derivation 과정에서 패킹된 프레임의 최소값과 최대값을 이용하여 양자화 범위를 설정한다. 다음으로, normalization을 수행하여 패킹된 feature를 0에서 1 사이의 값으로 스케일링하여 정규화된 feature y_{np}를 생성한다. 이는 원본 feature와 동일한 모양을 유지하면서도 정규화된 값을 가지도록 조정하는 과정이다. 마지막으로, quantization을 통해 내부 코덱 기준에 따라 32비트 부동 소수점에서 10비트 정수로 변환한다. 이 과정에서 양자화된 feature y_q는 정규화된 feature와 동일한 형태를 유지한다.

FCTM의 내부 코덱은 intra와 inter coding을 위해 VVC(Versatile Video Coding)를 사용한다. VVC는 이미지 데이터셋을 위해 하나의 양자화된 feature 프레임으로 구성된 YUV 형식으로 저장된 양자화 feature y_q를 입력으로 받는다. 휘도와 색차 성분으로 밝기와 색 정보를 분리하는 표준 영상 포맷으로 저장된 YUV는 VVC Encoder에 의해 single-bitstream b_x로 압축된다.

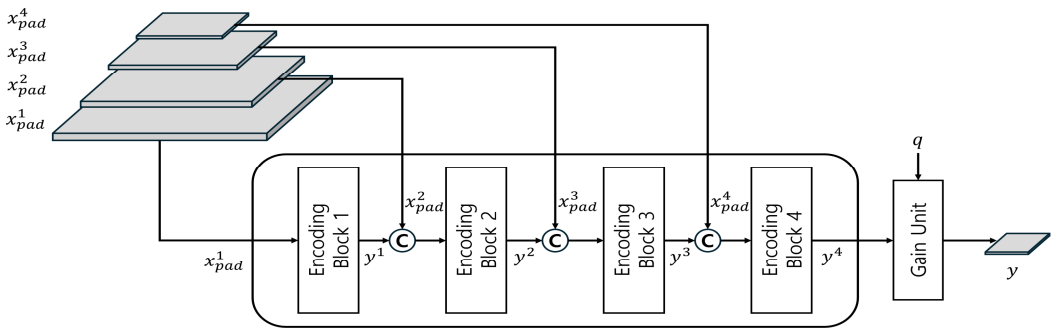

그림 2. FENet 아키텍처. 다양한 해상도의 패딩된 feature들이 각 인코딩 블록에 순차적으로 결합되며, 최종적으로 Gain Unit을 통해 출력 feature를 생성한다.

Fig. 2. Overall architecture of FENet. Padded features at multiple resolutions are progressively integrated through encoding blocks, and the final output is produced via a Gain Unit.

1.2 FCM Decoder

FCM Decoder는 Encoder의 역순으로 Inner Codec, Inverse Feature Conversion, Feature Restoration 순으로 진행된다. FCM Encoder에서 추출된 single-bitstream b_x는 VVC Decoder에 의해 디코딩되어 재구성된 YUV가 생성된다. 복원된 양자화 feature y_q는 재구성된 YUV에서 생성된다.

Inverse Feature Conversion에서는 내부 코덱에서 디코딩된 feature y_q를 입력으로 받아 3차원 feature 텐서로 변환하는 과정으로, dequantization, unpacking 순으로 진행된다. Feature Restoration에서는 feature inverse transform이 수행되며, 이 과정에서 그림 3에서 제시하는 DRNet(The Feature Decoding and Reconstruction Network)이 사용된다. DRNet은 압축 또는 변환된 feature를 정교하게 복원하기 위한 네트워크이다.

먼저 $\hat{y}$의 각 채널에 inverse gain unit에서 최고 품질에 대한 inverse gain vector 각 값을 곱한다. 그 후, inverse unit($\hat{z}_1$)의 출력은 디코딩 블록으로 유도되어 최고 해상도 $\hat{x}_{up}^1$의 중간 feature를 복원한다. 또한, 나머지 중간 feature들을 복원하기 위해 각 중간 feature인 $\hat{x}_{up}^2$에서 $\hat{x}_{up}^N$에 해당하는 디코딩 블록이 feature mixing block에 의해 사용된다. Feature mixing block은 해당 디코딩 블록의 출력과 복원된 하위 레이어 feature인 $\hat{x}_{up}^{N-1}$을 추가 입력으로 통합하여 중간 feature인 $\hat{x}_{up}^N$의 재구성 품질을 향상시킨다.

2. AAC

AAC는 MPEG-2와 MPEG-4 오디오 코덱 기술로, 인간 청각 특성을 활용해 데이터 용량을 줄이면서도 고음질을 제공하는 오디오 압축 방식이다[8]. 다양한 응용 시나리오에 맞춰 여러 프로파일이 사용되며, 주요 프로파일은 다음과 같다.

AAC-LC(Low Complexity): 가장 널리 사용되는 프로파

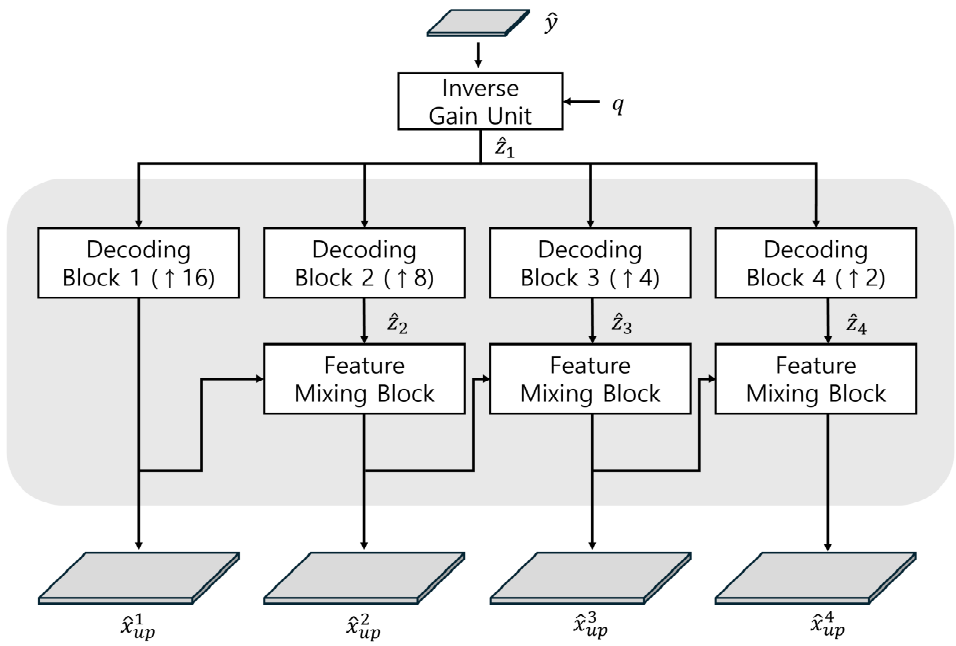

그림 3. DRNet 아키텍처. Inverse Gain Unit 출력은 Decoding Block과 Feature Mixing Block을 통해 순차적으로 처리되며, 각 단계에서 하위 해상도 feature를 통합하여 중간 feature를 재구성한다.
Fig. 3. DRNet architecture. The output from the Inverse Gain Unit is progressively decoded through Decoding and Feature Mixing Blocks, reconstructing intermediate features using lower-resolution inputs.

일로, 스테레오 신호에 대해 128kbp 이상의 비트레이트에서 원음과 구별하기 힘든 투명한 음질을 제공한다. 이 프로파일은 스트리밍 서비스와 방송과 같은 고음질이 요구되는 환경에서 주로 사용된다.

HE-AAC(High Efficiency AAC): 고주파수 대역을 저주파수 대역 성분과 적은 양의 부가 정보만으로 복원하는 SBR(Spectral Band Replicaton) 기술을 활용해 AAC-LC보다 낮은 비트레이트에서도 우수한 음질을 제공한다. HE-AAC는 스테레오 신호를 48kbps 내외로 압축할 수 있으며, 모바일 스트리밍 환경에서 자주 사용된다.

HE-AAC v2: HE-AAC를 확장하여 Parametric Stereo (PS) 기술을 추가로 적용한 프로파일로, 스테레오 신호를 32kbps 내외에서도 적정한 음질을 제공할 수 있다. 이는 저대역폭 환경에서도 우수한 음질을 유지하도록 설계되었다.

xHE-AAC(Extended HE-AAC): HE-AAC v2의 기술에 추가적인 압축 도구를 결합하여 비트레이트를 더욱 낮추면서도 효율성을 극대화한 프로파일이다. 이는 IoT 디바이스 및 저전력 오디오 응용 프로그램과 같은 제한된 대역폭 환경에 적합하다.

본 논문에서의 오디오 입력 신호는 모노 채널, 16비트 양자화, 16kHz 샘플링 주파수를 사용하며, 국제표준으로 채택되어 방송·스트리밍 등에서 표준으로 쓰이고, 저비트레이트에서도 우수한 음질과 압축 효율이 검증된 AAC-LC 프로파일을 비교 기준으로 선정하였다. AAC-LC 프로파일은 32kbps 내외에서 원본과 유사한 음질을 제공함이 보고되어 있다[9].

3. Anomalous Sound Detection

DCASE(Detection and Classification of Acoustic Scenes and Events) 커뮤니티는 음향 데이터 분석 기술 발전을 위해 알고리즘 평가 워크숍과 챌린지를 개최하며, 다양한 데이터셋으로 기술 한계를 탐구하고 새로운 모델 개발을 지원한다.

ASD(Anomalous Sound Detection)는 기계에서 발생하는 소리가 정상(정상 동작)인지 비정상(잠재적 고장)인지 판별하는 작업으로, 산업 현장에서 조기 유지보수를 위해 필수적인 기술이다. 그러나 실무 환경에서는 비정상 데이터가 매우 드물고 라벨링이 어려운 경우가 많아, 정상 데이터만으로 학습을 진행하고 이상 여부를 판별하는 비지도 또는 준지도 학습 기반 ASD 방식이 주로 연구되고 있다.

이러한 배경에서 DCASE 2023 Task 2는 추가적인 데이터 수집이나 하이퍼파라미터 튜닝 없이, 단일 섹션의 정상 데이터만으로 ASD 시스템을 구축하는 First-Shot Unsupervised ASD 과제를 제안하였다. 여기서 '이상 탐지(anomaly detection)'는 정상 음향 패턴을 학습한 모델이 통계적으로 유의미하게 벗어나는 비정상 음향을 식별하는 과정을 의미한다. 해당 Task에서는 Auto-Encoder(AE) 기반의 Simple AE와 Selective Mahalanobis 방법론이 Baseline 모델로 제시되었다[10].

Simple AE는 Mel-스펙트로그램을 입력받아 재구성 시 발생하는 MSE(Mean Squared Error)를 최소화하며 학습된다. 산출된 MSE가 정상이 아닌 이상 점수로 사용되며, 값이 클수록 이상 가능성이 높다. Selective Mahalanobis는 동일한 입력을 활용하여 AE의 Latent Space에서 추출한 평균 벡터와 공분산 행렬을 기반으로 Mahalanobis 거리를 산출하고 이상 점수를 계산한다. 이 방식은 특히 새로운 기계 유형의 교차 도메인 일반화 성능을 개선하며, 거리가 클수록 이상 데이터일 확률이 높음을 의미한다.

그림 4는 본 연구에서 채택한 DCASE Baseline에서 제시하는 anomaly detection 파이프라인을 보여준다. Data-Loader는 전처리된 Mel-스펙트로그램을 불러와 AE에 입력한다. AE의 Encoder 부분은 데이터를 Latent Space로 압축하고, Decoder 부분은 이를 Reconstruction 단계에서 원래 차원으로 복원한다. 학습 시에는 입력과 복원 결과의 차이를 MSE Loss로 계산해 가중치를 최적화한다. 테스트 단계에서는 Latent Space 벡터로부터 도메인별 공분산 행렬을 이용해 Mahalanobis Distance를 산출하거나, Reconstruction과 입력 간 MSE Loss를 그대로 활용해 이상도 점수를 얻는다. 해당 점수가 임계값을 넘으면 Anomaly Detection 블록이 샘플을 이상으로 판정하며, 최종 결과는 Performance Evaluation에서 평가 지표(F1 score 등)로 정량화된다.

본 논문에서는 DCASE Task 2를 채택하여 Baseline 모델에 대해 다음 두 가지 입력 방식을 적용하여 실험을 진행하였다. 첫째, 원본 WAV를 AAC로 압축·복원한 후 Mel-스펙

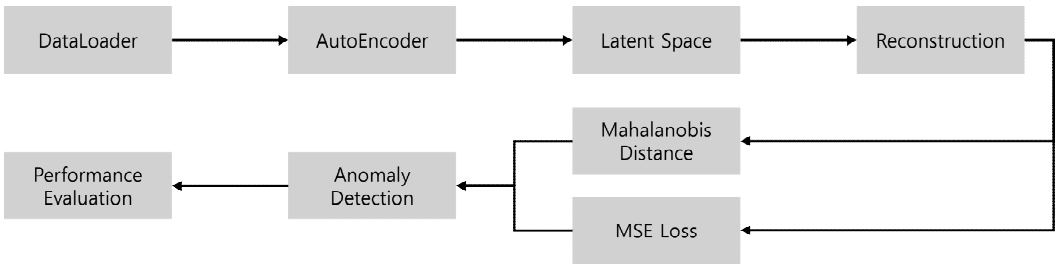

그림 4. DCASE Baseline 구조. AE로부터 추출된 Latent Space 및 복원 결과를 기반으로 MSE Loss와 Mahalanobis Distance를 계산하여 이상 여부를 판단하고, 성능을 평가한다.
Fig. 4. DCASE Baseline architecture. Anomaly detection is performed based on MSE loss and Mahalanobis distance computed from the latent space and reconstruction output of the AE, followed by performance evaluation.

트로그램 이미지를 YUV 4:0:0 포맷으로 변환하여 입력하였고, 둘째, 동일한 스펙트로그램 이미지를 FCM 인코딩-디코딩을 거쳐 복원된 feature tensor 형태로 변환하여 입력하였다.

Ⅲ. 제안 방법

본 논문은 1차원 오디오 신호를 스펙트로그램으로 변환한 뒤, FCM 기반 ACoM을 이용하여 핵심 feature map을 추출·압축하고, 복원된 feature tensor를 이상 음향 탐지를 위한 입력으로 활용하는 구조를 제안한다.

1. 압축 방식

1.1 ACoM

그림 5는 ACoM의 전체 네트워크 구조를 나타낸다. ACoM Encoder에서는, 입력된 오디오 신호를 먼저

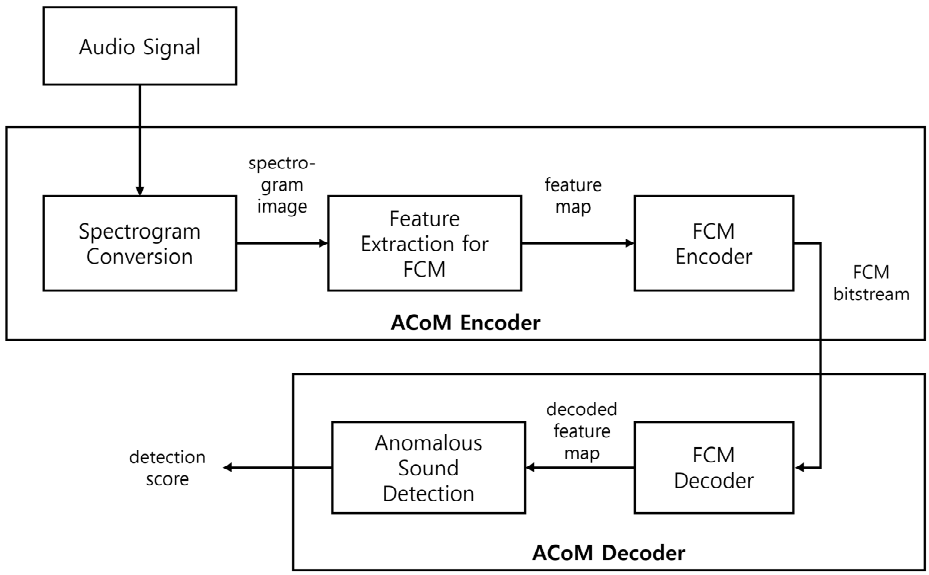

그림 5. ACoM 아키텍처 개요. Mel-스펙트로그램에서 추출된 feature map을 FCM으로 압축 및 복원한 후, 이상 탐지에 활용한다.
Fig. 5. Overview of the ACoM architecture. Feature maps extracted from Mel-spectrograms are compressed and reconstructed via FCM, then used for anomaly detection.

Spectrogram Conversion 모듈을 통해 Mel-스펙트로그램으로 변환하고, 이를 기반으로 Feature Extraction for FCM 모듈에서 이상 탐지에 유용한 feature map을 추출한다. 추출된 feature map은 FCM Encoder를 통해 압축되어 FCM bitstream으로 생성된다. 이 과정은 전송 또는 저장 효율을 극대화하기 위한 단계로, FCM Encoder의 내부 압축 방식으로는 VVC 기반 VTM 12.0이 사용되며, QP(Quantization Parameter)를 20에서 35까지 조정하여 압축률이 높아질수록 이상 탐지 성능이 얼마나 유지되는지를 분석할 수 있도록 구성하였다.

ACoM Decoder에서는, 수신된 FCM bitstream을 FCM Decoder가 복원하여 decoded feature map을 생성한다. 이 복원된 feature는 Anomalous Sound Detection 모듈에 입력되어, 정상 여부를 판단하고 detection score를 출력하게 된다.

기존 FCM은 객체 검출을 위한 이미지 feature map을 대상으로 학습된 모델이다. 본 연구에서는 별도의 재학습 없이, 해당 영상용 FCM 파라미터를 그대로 적용하여 Mel-스펙트로그램 feature를 압축·복원하였다. 따라서 ACoM은 추가 학습 비용 없이도 anomaly detection에 유용한 feature만을 선택적으로 압축하여 불필요한 데이터를 줄이고, 낮은 비트레이트 환경에서도 높은 탐지 성능을 유지함으로써 제한된 대역폭·저장 자원에서도 지연을 최소화할 수 있다.

1.2 AAC

AAC 압축 프로필 중 가장 범용성이 높은 AAC-LC를 채택하여 압축 실험을 설계하였다. 그림 6은 AAC 기반 실험의 전체 흐름을 시각적으로 나타낸다. 먼저, 원본 WAV 오디오가 AAC Encoder로 입력되며, 이 과정에서 인간 청각 특성을 반영한 psychoacoustic 모델이 적용되어 청각적으로 덜 중요한 주파수 대역의 정보가 제거된다. 또한, 오디오 신호는 AAC Encoder에서 4~32kbps 사이의 다양한 비트레이트로 압축되어, 압축 수준에 따른 성능 변화를 비교, 평가할 수 있도록 하였다. 이후 AAC Decoder를 통해 압축된 오디오가 복원되고, 복원된 신호에 대해 STFT(Short-time Fourier transform)를 수행하여 2차원 스펙트로그램을 생성한다. 이렇게 생성된 스펙트로그램을 ASD에 필요한 입력으로 사용하며 다양한 비트레이트별 ASD 모델 성능을 확인한다.

2. Training

본 연구에서는 DCASE 2024 Task 2 데이터셋 중 Fan, Slider, ToyCar, Valve 등 네 가지 기계 소리를 선정하여 실험을 진행하였다[1]. 각 기계 소리별 데이터셋은 학습(Training)에 사용되는 정상 데이터 1,000개와 테스트(Test)에 사용되는 정상 100개 및 비정상 100개, 총 200개로 구성되어 있으며, 모든 WAV 파일은 16kHz, 16-bit PCM 규격의 단일 채널(모노) 신호로 약 10초 분량이다.

그림 7은 Fan, Slider, ToyCar, Valve 네 기계 소리의 정상(normal)과 비정상(anomaly) Mel-스펙트로그램 예시를 제시한다. 각 행은 기계 종류를, 좌측 열은 정상 신호를, 우측 열은 비정상 신호를 나타낸다. 정상 스펙트로그램은 전 대역에 걸쳐 고른 에너지 분포와 일정한 주파수 패턴을 보이는 반면, 비정상 스펙트로그램에서는 특정 주파수 대역의 진폭 급등·수평 스트라이프(회전 불균형)·불규칙 잡음 패턴 등이 두드러져 기계 이상을 시사한다. 이러한 시각적 차이는 이후 AE 기반 ASD 모델이 이상 여부를 판단할 핵심 단서로 활용된다.

본 실험에서는 각 오디오 신호를 별도의 리샘플링 없이 Mel-스펙트로그램으로 변환한 뒤, 압축 방식에 따라 두 가지 입력 경로로 분기하였다. AAC 경로에서는 변환된 Mel-스펙트로그램을 WAV로 재합성한 후, AAC-LC 코덱으로 압축하고 복원하였다. 복원된 오디오는 다시 Mel-스펙트로

그림 6. AAC 기반 ASD 파이프라인. WAV 오디오는 AAC-LC로 압축·복원된 후 스펙트로그램으로 변환되어 ASD 입력으로 사용된다.
Fig. 6. AAC-based ASD pipeline. WAV audio is compressed and decoded using AAC-LC, then converted to a spectrogram for ASD input.

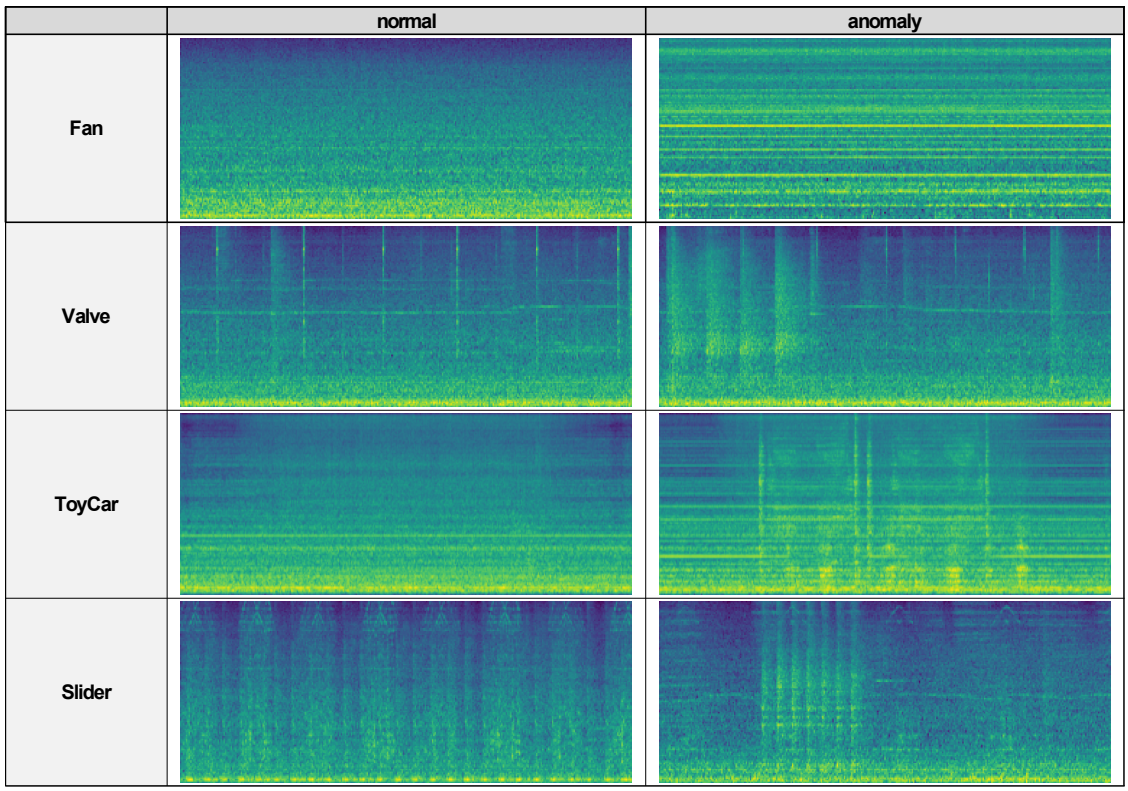

그림 7. 각 기계별로 입력되는 normal, anomaly 데이터셋에 대한 예시
Fig. 7. Example inputs of the normal and anomaly datasets for each machine

그램 이미지로 변환되어 ASD의 입력으로 사용되며, 모델은 이미지 기반 입력을 압축·복원하는 구조로 학습된다.

FCM 경로에서는 Mel-스펙트로그램으로부터 CNN 기반 네트워크를 통해 320차원 feature map을 추출하고, 이를 FCM 코덱으로 압축·복원하여 생성된 feature tensor를 ASD에 입력하였다. 이 feature tensor 기반 입력은 기계가 필요한 정보를 중심으로 학습하도록 설계되었다.

두 압축 경로에서 생성된 데이터는 서로 다른 통계적 특성을 가지므로, 각각 독립된 AE 기반 ASD 모델을 별도로 학습시켜 입력 방식별 최적화를 유도하였다. AE는 입력 데이터를 저차원 latent space로 압축한 뒤 다시 복원하는 과정을 통해 학습되며, 이때 Batch Normalization과 ReLU 활성화 함수를 적용하여 안정적인 학습을 유도한다. 학습 완료 후 latent space에서 얻은 평균 벡터와 공분산 행렬을 바탕으로 입력 데이터의 latent vector에 대한 Mahalanobis Distance를 계산하고, 거리가 클수록 이상 음향일 가능성이 높다고 판단한다. 최종적으로 F1-score 지표를 통해 압축 방식별 이상 음향 탐지 성능을 정량적으로 검증한다.

Ⅳ. 실험 결과

본 장에서는 DCASE 2024 Task 2가 first-shot unsupervised ASD 과제로, 정상 데이터로 한 번 학습한 모델을 그대로 모든 장비에 적용해야 한다는 특성을 따른다. 이때 장비마다 음향 주파수 대역·주기성·잡음 환경이 달라 재구성 오차 분포가 다르게 나타나므로 Fan, Slider, ToyCar, Valve 네 기계를 각각 따로 평가하였으며, 성능 평가 지표로는 Precision과 Recall에서 파생된 F1-Score를 사용하였다. 이에 따라 ACoM과 AAC 두 압축 기법의 성능을 비교한 결과를 표 1에 정리하였다. 표에는 각 데이터셋에서 도출된 비트레이트(bitrate, kbps)와 F1 score가 함께 제시되어

표 1. 각 기계 유형별 F1-score 및 사용된 kbps
Table 1. F1-score and kbps used for each machine type

Dataset	Raw	AAC		ACoM(proposal)		
	F1 score	kbps	F1 score	QP	kbps	F1 score
Fan	0.675	32	0.645	20	6.3	0.667
		24	0.662	25	3.3	0.664
		16	0.638	30	1.7	0.660
		4	0.638	33	1.3	0.655
Valve	0.665	32	0.667	20	6.3	0.667
		24	0.658	25	3.3	0.654
		16	0.667	30	1.7	0.592
		4	0.641	33	1.3	0.571
ToyCar	0.662	32	0.622	20	6.3	0.629
		24	0.524	25	3.3	0.624
		16	0.632	30	1.7	0.641
		4	0.594	33	1.3	0.605
Slider	0.644	32	0.667	25	3.3	0.601
		24	0.521	30	1.7	0.569
		16	0.581	33	1.3	0.469
		8	0.667	35	0.9	0.465

두 방식의 전반적인 성능을 한눈에 확인할 수 있다. 그림 8-11은 이러한 결과를 시각적으로 나타낸 그래프로, 가로축에 비트레이트, 세로축에 F1 score를 배치하였다. ACoM은 낮은 비트레이트에서도, 더 높은 비트율로 압축된 AAC와 유사한 성능을 유지함을 확인할 수 있었다.

Fan 유형에서의 성능을 비교한 결과, ACoM은 전반적으로 F1-score가 0.65 이상을 유지하며 낮은 비트레이트에서도 안정적인 성능을 보였다. 특히, ACoM은 1.3kbps의 극히 낮은 비트레이트에서도 높은 성능을 기록했으며, 이는 4kbps에서의 AAC보다 더 우수한 결과였다. 또한, ACoM

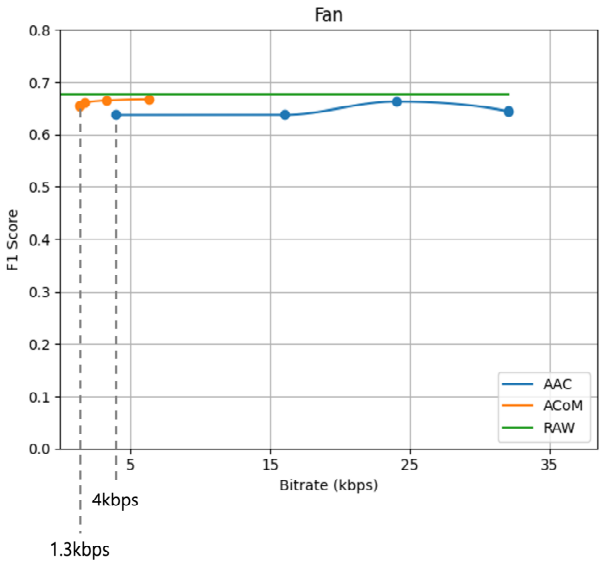

그림 8. Fan의 실험 결과 그래프
Fig. 8. Experimental results graph for the Fan

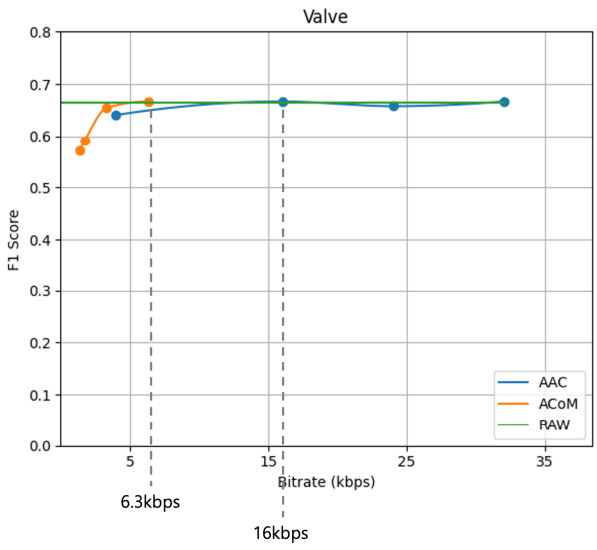

그림 9. Valve의 실험 결과 그래프
Fig. 9. Experimental results graph for the Valve

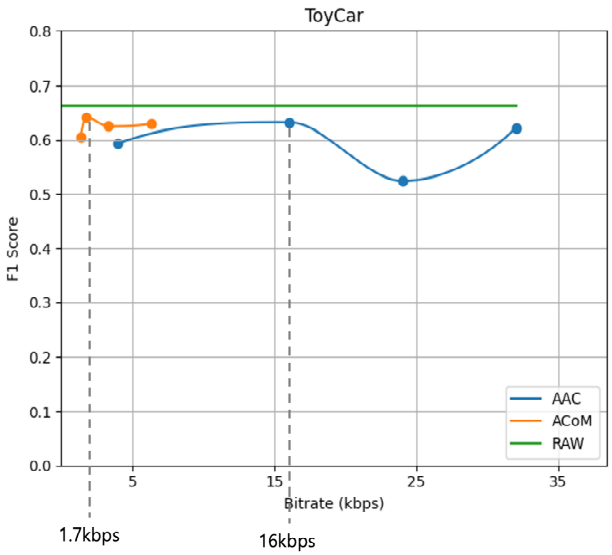

그림 10. ToyCar의 실험 결과 그래프
Fig. 10. Experimental results graph for the ToyCar

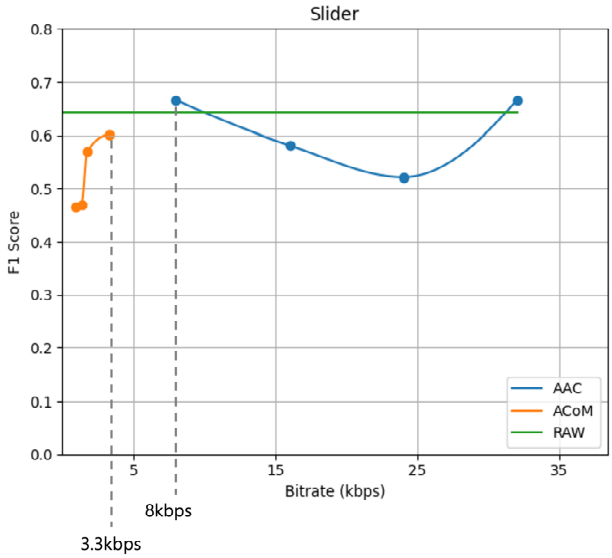

그림 11. Slider의 실험 결과 그래프
Fig. 11. Experimental results graph for the Slider

은 비트레이트 증가에 따른 성능 변동이 크지 않아 압축률이 높아도 성능 저하가 최소화되는 경향을 보였다. 이는 ACoM이 효과적인 feature 압축 방식을 활용하여 중요한 정보를 보존하는 데 강점이 있음을 시사한다. 반면, AAC는 4kbps에서는 ACoM보다 낮은 성능을 보였으나, 16~32 kbps 범위에서는 비교적 안정적인 성능을 유지했다. 하지만 32kbps 이상의 높은 비트레이트에서는 성능이 다소 감소하는 현상이 관찰되었다. 이는 모노 오디오 신호를 부호화 하는데 16kbps 내외가 적절한 비트레이트이며, 32kbps 이상에서는 더 이상의 성능 개선은 발견되지 못함을 시사한다. 결과적으로, Fan 유형에서는 ACoM이 낮은 비트레이트에서도 성능 저하 없이 안정적인 결과를 제공하기 때문에, 제한된 비트레이트 환경에서도 효과적인 활용이 가능할 것으로 보인다.

Valve 유형에서의 성능을 비교한 결과, 낮은 비트레이트 구간에서는 ACoM이 AAC보다 다소 낮은 F1-score를 기록했으나, 특정 성능 기준에서의 효율성을 비교했을 때 ACoM이 더 우수한 결과를 보였다. 즉, F1-score가 0.667에 도달하는 시점을 살펴보면, ACoM은 6.3kbps에서 이 성능을 달성한 반면, AAC는 16kbps의 더 높은 비트레이트가 필요했다. 이는 ACoM이 AAC 대비 절반 이하의 데이터량으로 유사한 성능을 유지할 수 있음을 의미하며, 데이터 전송이나 저장 용량이 제한된 환경에서 보다 효율적인 선택이 될 수 있음을 시사한다.

또한, 그래프에서 확인할 수 있듯이 AAC는 비트레이트가 증가할수록 성능이 점진적으로 향상되는 경향을 보이며, ACoM과의 성능 격차가 점차 줄어드는 모습을 보인다. 하지만 낮은 비트레이트 구간에서는 ACoM이 안정적인 성능을 유지하며, 상대적으로 적은 데이터량으로 높은 성능을 제공하는 것이 특징이다.

따라서, Valve 유형에서는 낮은 비트레이트 환경에서 ACoM이 AAC보다 더 효율적인 데이터 압축 성능을 제공하며, 제한된 네트워크 대역폭이나 저장 공간에서 유리하게 활용될 수 있을 것으로 판단된다.

ToyCar 유형에서의 성능을 비교한 결과, ACoM은 전반적으로 F1-score가 0.6 이상을 유지하며 안정적인 성능을 보여주었다. 특히, 1.7kbps의 매우 낮은 비트레이트에서 F1-score가 0.641을 기록하였으며, 이는 실험된 모든 비트레이트 중 가장 높은 성능이다. 이는 ACoM이 낮은 비트레이트에서도 효과적으로 중요한 음향 특징을 보존하며, 높은 데이터 압축률을 유지하면서도 성능 저하를 최소화하는 능력을 갖추고 있음을 시사한다. 반면, AAC는 비트레이트

가 증가함에 따라 성능이 안정적으로 증가하는 것이 아니라, 특정 비트레이트 구간(약 15~25kbps)에서 F1-score가 급격히 감소하는 경향을 보였다. 이는 해당 구간에서 AAC의 압축 방식이 중요한 특징을 충분히 유지하지 못하거나, 특정 비트레이트에서 데이터 손실이 크게 발생했기 때문으로 추정된다.

이러한 결과를 종합적으로 고려했을 때, ToyCar 유형에서는 ACoM이 낮은 비트레이트에서도 성능 저하 없이 안정적인 결과를 보이므로, 데이터 전송 효율성과 성능을 동시에 고려해야 하는 환경에서 더욱 적합한 선택이 될 것으로 판단된다.

Slider 유형에서의 성능을 비교한 결과, ACoM은 비트레이트가 증가함에 따라 점진적으로 F1-score가 향상되는 이상적인 경향을 보였다. 그러나 전반적인 성능을 보면 ACoM은 AAC보다 낮은 F1-score를 기록하였다. 특히, ACoM의 F1-score는 최대 약 0.55 수준에서 머무른 반면, AAC는 일부 비트레이트 구간에서 0.7에 가까운 높은 성능을 보였다. 이러한 차이는 AAC가 특정 비트레이트에서 더 효과적으로 음향 정보를 보존하는 반면, ACoM은 아직 FCM이 충분히 학습되지 않은 상태에서 적용되어 중요한 특징을 최적으로 추출·복원하지 못했기 때문으로 보인다. 또한, AAC는 낮은 비트레이트에서는 비교적 높은 성능을 보이지만, 10~25kbps 구간에서는 성능이 급격히 감소하는 패턴을 보였다. 반면 ACoM은 낮은 비트레이트에서도 비교적 안정적인 성능을 유지하는 경향을 보였다.

Slider의 경우, normal과 anomaly 간 스펙트로그램 차이가 미세하고 단조로운 특성을 보여 모델이 이상 여부를 구분하기 어려운 점도 성능 저하의 한 원인으로 판단된다. 따라서 Slider 유형에서는 AAC가 높은 비트레이트에서 더 나은 성능을 발휘하지만, FCM 학습이 충분히 이루어진다면 ACoM 역시 안정적인 대안이 될 수 있을 것으로 기대된다.

Ⅴ. 결 론

본 연구에서는 FCM 압축 방식을 기반으로 한 새로운 오디오 압축 기법인 ACoM을 제안하고, 이를 ASD에 적용하여 성능을 검증하였다. 실험 결과, ACoM을 활용한 경우 전반적인 F1-score가 60~70%에 머무르는 것을 확인했다. 이는 이진 분류 방식의 한계로 인해 세밀한 이상 음향 분류가 어려웠던 결과로 분석된다. 그럼에도 불구하고, 기존 오디오 코덱인 AAC와의 비교 실험에서 ACoM은 낮은 비트레이트 환경에서도 유사한 수준의 탐지 성능을 유지하였으며, 비트레이트 증가에 따라 점진적인 성능 향상을 보이는 이상적인 패턴이 관찰되었다. 반면, AAC는 비트레이트 변화에 따른 성능 편차가 상대적으로 불안정한 모습을 나타냈다.

산업 현장이나 공장에서 활용 가능한 수준의 90% 이상의 성능을 확보하기 위해서는 성능을 개선할 필요가 있다. 이를 위해 기존 FCM이 객체 탐지를 위한 task에 맞추어 학습되어 있어, 오디오 기반 task에 적합하도록 학습 코드를 수정한 후 재훈련할 계획이다. 또한, anomaly detection 뿐만 아니라 보다 세밀한 성능 평가가 가능한 task 등을 통해 실험을 진행할 예정이며, 현재 약 1,000개의 제한된 트레이닝 데이터를 사용함에 따른 과적합 위험을 해소하기 위해 다양한 데이터셋을 확보하고 적용할 계획이다.

결론적으로, 본 연구는 FCM 기반 ACoM이 낮은 비트레이트 환경에서도 유의미한 anomaly detection 성능을 확보할 수 있음을 실험적으로 입증하였으며, 향후 FCM 학습 방식의 개선과 대규모 데이터셋 및 다양한 task 적용을 통해 ACoM의 효율성과 정확도를 한층 강화함으로써, 다양한 산업 환경에서 활용 가능한 차세대 이상 음향 탐지 모델을 발전시킬 수 있을 것으로 기대된다. 또한, 본 연구는 ACoM이 기계를 위한 오디오 코덱의 표준으로 자리잡을 수 있는 가능성 또한 제시한다.

참 고 문 헌 (References)

[1] ISO/IEC JTC I/SC 29/WG 06 MPEG Audio Coding, “Use Cases and Requirements on Audio Coding for Machines,” in Proceedings of the 149th Meeting of the Moving Picture Experts Group (MPEG), Geneva, Jan. 2025.

[2] H. Choo, W. Cheong, and J. Seo, “*Standardization Trends in Video Coding for Machines,*” Broadcasting and Media Magazine, Vol.28, No.1, pp.38-52, 2023.

[3] S. Jeon, D. Lee, H. Choo, and J. Seo, “Performance Improvement of FCM using End-To-End Deep Learning-Based Codec,” Proceedings of the KIBME (Korean Institute of Broadcasting and Media Engineers) Conference, Jeju, Jun. 2024.
[4] H. Han, S. Lee, S. Jung, J. Kim, S. Kwak, J. Lee, W. Cheong, H. Choo, and H. Choi, “Multi-Scale Feature Compression Method based on MSE Loss for FCM,” Proceedings of the KIBME (Korean Institute of Broadcasting and Media Engineers) Conference, Jeju, Jun. 2024.
[5] H. Jeong, S. Jang, D. Lim, and H. Kim, “Normalized Fused Feature Map Encoding for FCM,” Proceedings of the KIBME (Korean Institute of Broadcasting and Media Engineers) Conference, Jeju, Jun. 2024.
[6] S. Byun and J. Seo, “Analysing Trends in Audio Coding for Machines,” Proceedings of the KIBME (Korean Institute of Broadcasting and Media Engineers) Conference, Jeju, Jun. 2024.
[7] ISO/IEC JTC I/SC 29/WG 04 MPEG Video Coding, “Algorithm Description of FCTM,” in Proceedings of the 148th Meeting of the Moving Picture Experts Group (MPEG), Kemer, Nov. 2024.
[8] ISO/IEC JTC I/SC 29/WG 11, “The MPEG AAC Family of Codecs,” in Proceedings of the 126th Meeting of the Moving Picture Experts Group (MPEG), Geneva, Mar. 2019.
[9] M. Matousek, J. Schimmel, and J. Vondra, “*The influence of the bitrate level on the subjective sound quality perception of the concatenated non-entropic audio coding algorithms in the digital broadcasting chain,*” Radioengineering, Vol. 29, No. 4, pp. 672 – 682, Dec. 2020. doi: https://doi.org/10.13164/re.2020.0672
[10] T. Nishida, et al., “Description and discussion on DCASE 2024 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring,” arXiv preprint arXiv:2406.07250, 2024.
[11] T. Nishida, et al., “DCASE 2024 Challenge Task 2 Development Dataset,” Zenodo, 2024. doi: https://doi.org/10.5281/zenodo.10902294.

저 자 소 개

변 수 빈

- 2022년 2월 : 동아대학교 컴퓨터공학과(컴퓨터공학사)
- 2024년 ~ 현재 : 동아대학교 컴퓨터공학과 석사과정
- ORCID : https://orcid.or/0009-0005-2662-5302
- 주관심분야 : 멀티미디어, 오디오/비디오 부호화, 딥러닝, 머신 비전

서 정 일

- 2005년 2월 : 경북대학교 전자공학과(공학박사)
- 1998년 ~ 2000년 : LG반도체 선임연구원
- 2000년 ~ 2023년 : 한국전자통신연구원 실감미디어연구실장
- 2023년 ~ 현재 : 동아대학교 컴퓨터공학과 부교수
- ORCID : https://orcid.or/0000-0001-5131-0939
- 주관심분야 : 멀티미디어, 오디오/비디오 부호화, 딥러닝, 머신 비전