특집논문 (Special Paper)
방송공학회논문지 제30권 제3호, 2025년 5월 (JBE Vol.30, No.3, May 2025)
https://doi.org/10.5909/JBE.2025.30.3.279
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

기계를 위한 오디오 부호화 표준화 동향 분석

임 우 택[a], 장 인 선[a)‡], 백 승 권[a)], 강 정 원[a)]

Standardization Trends in Audio Coding for Machines

Wootaek Lim[a)], Inseon Jang[a)‡], Seungkwon Beack[a)], and Jungwon Kang[a)]

요 약

최근 인공지능과 머신러닝 기반의 오디오 처리 기술이 급속히 발전함에 따라, 인간 청취 품질을 극대화해온 기존의 오디오 코딩 방식과 달리 기계 학습 및 분석에 최적화된 새로운 접근 방식이 요구되고 있다. 이에 대응하여 MPEG에서는 기계를 위한 오디오 부호화(Audio Coding for Machines, ACoM) 표준화를 추진 중이며, 다양한 사용 사례와 기술 요구 사항을 도출하고 있다. 본 논문에서는 기존 오디오 코딩 기술의 한계, ACoM 기술의 개념과 목표, 범위와 시스템 개요, 주요 사용 사례, 표준 요구 사항 등을 살펴보고, MPEG 내에서 진행되는 표준화 동향과 향후 로드맵을 전망한다. 또한 ACoM 기술이 오디오 분석, 음성 인식, 스마트 디바이스 및 IoT 환경에서 어떠한 역할을 수행할 수 있는지 조망하며, 이를 위해 해결해야 하는 기술적 과제 및 향후 표준화가 산업 및 기술 발전에 미칠 수 있는 영향에 대해 논의한다.

Abstract

With the rapid advancement of artificial intelligence and machine learning-based audio processing technologies, a new approach optimized for machine learning and analysis has become necessary, as opposed to conventional audio coding methods that are primarily aimed at maximizing human listening quality. In response, MPEG is actively pursuing the standardization of Audio Coding for Machines (ACoM), deriving various use cases and technical requirements. This paper examines the limitations of conventional audio coding methods and introduces the concepts, objectives, scope and system overview of ACoM technologies. It further outlines key use cases and standardization requirements, and presents ongoing standardization activities within MPEG, as well as an outlook on future roadmaps. Additionally, this paper discusses potential roles of ACoM technologies in audio analysis, speech recognition, smart devices, and IoT environments. Finally, we discuss the technical challenges that need to be addressed and the impact that future standards may have on industry and technological advancements.

Keyword : Audio coding for machines, Audio coding, Speech coding, Machine learning, Artificial intelligence

a) 한국전자통신연구원 미디어부호화연구실(Electronics and Telecommunications Research Institute, Media Coding Research Section)
‡ Corresponding Author : 장인선(Inseon Jang)
E-mail: jinsn@etri.re.kr
Tel: +82-42-860-5791
ORCID: https://orcid.org/0000-0003-2237-2668

※ 이 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. RS-2017-II170072, 초실감 테라미디어를 위한 AV 부호화 및 LF 미디어 원천기술 개발).
※ This work was supported by Institute for Information & communi- cations Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. RS-2017-II170072, Development of Audio/Video Coding and Light Field Media Fundamental Technologies for Ultra Realistic Tera-media).
· Manuscript March 26, 2025; Revised April 30, 2025; Accepted May 2, 2025.

Ⅰ. 서 론

디지털 혁신이 가속화됨에 따라 인공지능(AI)과 사물인터넷(IoT) 기기의 활용이 급증하고 있으며, 스마트 팩토리, 스마트홈, 자율주행, 의료 모니터링과 같은 다양한 산업 분야에서 기계가 직접 오디오 데이터를 처리하고 활용하는 사례가 증가하고 있다. 이러한 환경에서는 예측 유지보수(Predictive maintenance), 품질 관리, 에너지 최적화 및 치안 감시 등의 다양한 목적으로 방대한 오디오 데이터가 지속적으로 수집되고 분석된다. 그러나 기존 오디오 코딩 기술은 이러한 초다채널 및 초광대역 오디오 데이터를 실시간으로 효율적으로 처리하는 데 한계가 있다. 특히, IoT 기기의 확산으로 인해 기계가 소비하는 오디오 데이터의 규모는 지속적으로 증가하고 있으며, 2023년 157억 개였던 IoT 기기 연결 수는 2029년 389억 개에 이를 것으로 예상된다[1]. 이러한 변화 속에서 기존 오디오 코딩 기술의 인간의 청취 경험을 최적화하는 방식에서 벗어나 기계가 보다 효율적으로 오디오 데이터를 소비할 수 있도록 하는 새로운 접근 방식이 요구된다. 이에 따라 기계를 위한 오디오 부호화(Audio Coding for Machines, ACoM) 기술의 필요성이 더욱 부각되며, 이를 통해 기계 학습 및 자동화 시스템의 효율성을 극대화할 수 있는 기술적 기반이 마련될 것으로 기대된다.

기존 오디오 코딩 기술은 심리음향 모델을 기반으로 인간의 청각 특성을 고려한 압축 방식을 사용하여 가청 주파수 대역(20Hz ~ 20kHz) 내에서 인지 품질을 유지하면서도 오디오 데이터 압축률을 극대화하는 것을 목표로 한다. 그러나 기계 중심 오디오 데이터는 인간의 청각 특성과 무관한 경우가 많으며, 오히려 비가청 대역(초음파 및 저주파 영역)의 정보를 포함한 대용량 데이터를 효과적으로 처리할 필요도 발생한다. 또한, 기존 오디오 코딩 방식은 주로 스테레오(2채널) 또는 서라운드(5.1채널) 환경을 기준으로 설계되었으나, 스마트 팩토리나 의료 모니터링과 같은 분야에서는 수십에서 수백 개에 이르는 센서를 활용한 다채널 오디오 데이터가 생성되므로, 기존 기술로는 이러한 데이터를 실시간으로 효율적으로 압축하고 처리하는 것이 어렵다.

이러한 한계를 극복하기 위해 AI와 머신러닝 기반의 새로운 오디오 코딩 기술이 개발되고 있다. 기존 신호처리 기반 오디오 압축 방식에서 벗어나 신경망 기반 오디오 코덱(Neural Audio Codec)이 등장하면서, 기계 학습 시스템에 최적화된 오디오 데이터 처리가 가능해졌다[2]. AI 기반 오디오 코딩 기술은 주파수와 시간 도메인 정보를 보존하면서도 압축률을 향상시켜, 기계가 오디오 데이터를 보다 정밀하게 분석하고 실시간으로 처리할 수 있도록 한다.

이와 유사하게, 기계가 분석하고 활용할 수 있는 비디오 데이터의 효율적인 부호화를 위한 연구도 활발히 진행되고 있다. Video Coding for Machines(VCM) 기술은 기존 비디오 압축 기술이 인간의 시각적 경험을 최적화하는 데 초점을 맞춘 것과 달리, 기계 학습 및 분석을 위한 최적화된 비디오 표현 방식을 제공한다. 또한, Feature Coding for Machines(FCM)은 기계가 분석하는 주요 특징만을 추출하여 압축하는 방식으로, 데이터 효율성을 극대화할 수 있도록 설계되었다[3]. 이러한 기계 중심 미디어 부호화 기술들은 다양한 산업 환경에서의 데이터 활용성을 높이기 위해 MPEG에서 표준화가 진행되고 있으며, 향후 AI 기반 데이터 처리의 핵심 기술로 자리 잡을 것으로 기대된다.

본 논문에서는 이러한 배경을 바탕으로 MPEG에서 진행 중인 ACoM 표준화 동향을 중심으로 기계 학습 및 기계 간 통신을 위한 오디오 코딩 기술의 발전 방향을 분석한다. 2장에서는 기존 오디오 코딩 기술의 발전 과정과 한계를 정리하고, 기계 중심 오디오 코딩의 필요성을 설명한다. 3장에서는 MPEG ACoM의 개념, 목표, 주요 사용 사례 및 표준화 진행 현황을 살펴보고, 토의 및 결론에서는 ACoM이 산업과 기술에 미칠 영향을 분석하고 향후 발전 방향을 제시한다. 본 논문을 통해, 기계 중심 오디오 코딩 기술이 AI 및 IoT 기반의 다양한 응용 분야에서 어떻게 활용될 수 있는지를 조망하고, MPEG ACoM 표준화가 미디어 기술의 새로운 패러다임을 형성하는 데 기여할 것임을 논의하고자 한다.

Ⅱ. ACoM을 위한 오디오 코딩 기술의 한계와 과제

오디오 코딩 기술은 수십 년간 발전하면서 음성 통신과

음악 스트리밍, 방송 등 다양한 분야에서 널리 활용되어 왔다. 특히 MPEG에서 표준화한 AAC(Advanced Audio Co-ding), USAC(Unified Speech and Audio Coding), MPEG-H 3D Audio 및 IETF에서 표준화한 Opus 등의 기술은 디지털 오디오 압축 및 전송 효율성을 높이는 데 크게 기여했다.

AAC는 MP3의 후속으로 MPEG-2와 MPEG-4 표준에서 발전된 오디오 코딩 기술이다[4]. AAC는 심리음향 모델을 적용하여 인간의 청각으로 인지되지 않는 정보들을 과감히 제거함으로써 동일한 비트레이트에서 MP3보다 더 높은 음질을 제공하였다. AAC는 음악 스트리밍 서비스 및 디지털 방송 등에서 널리 활용되고 있다. Opus는 낮은 비트레이트에서도 실시간 음성 및 음악 데이터 전송에 최적화된 코덱으로 인터넷 기반의 음성 통신 및 스트리밍 서비스에 주로 사용된다[5]. 특히 저지연 특성과 광범위한 음성 및 음악 신호 지원으로 실시간 통신에 특화된 특성을 갖추었다. USAC은 MPEG-D 표준으로 음성과 일반 음악 신호를 통합적으로 압축하는 최신 오디오 부호화 기술이다. 이 기술은 음성과 음악이 혼합된 콘텐츠에서도 안정적인 품질을 유지하면서 압축 효율을 높일 수 있도록 설계되었다[6]. 가장 최신 오디오 코딩 기술인 MPEG-H 3D Audio는 객체 기반 오디오와 공간음향 정보를 이용하여 몰입형 오디오 환경을 구현하는 차세대 오디오 표준이다. 사용자는 개인의 환경이나 취향에 따라 오디오 객체를 개별적으로 조정할 수 있어, 보다 현실감 있고 개인화된 청취 경험을 제공한다[7].

그러나 이러한 기존의 오디오 코딩 기술들은 대부분 인간의 청각적 특성을 중심으로 설계된 손실 압축 방식을 사용하고 있다. 심리음향 모델을 통해 인간이 잘 인지하지 못하는 정보를 제거하여 높은 압축률을 달성하지만, 이는 기계 분석 및 인공지능 기술을 활용한 분석 환경에서는 중요한 정보의 손실을 초래할 가능성이 크다. 예를 들어 음성 인식, 소리 이벤트 탐지, 기계 상태 모니터링 등의 응용에서는 인간이 인지하지 못하는 미세한 신호의 차이도 기계 분석의 정확도에 큰 영향을 미칠 수 있다. 특히 음성 부호화 기술의 경우 음성 신호만을 다루기 때문에 음성 이외의 다양한 환경음이나 복합적인 소리 신호의 전달에는 한계가 있다. 음성 신호에 특화된 LPC(Linear Prediction Coding)나 CELP(Code Excited Linear Prediction)와 같은 모델링 방식은 낮은 비트레이트에서 음성 전달 효율이 뛰어나지만, 일반적인 음향 정보를 처리할 때 품질이 급격히 저하되는 문제점이 있다[8,9].

ACoM은 인간 중심의 정보 전달 방식에서 벗어나 인공지능 기반의 기계 학습 및 신호 분석에 최적화된 오디오 부호화 기술을 목표로 한다. ACoM에서는 인간에게 불필요하거나 무의미하게 여겨진 정보도 기계의 관점에서는 중요한 분석 자료로 활용될 수 있다는 전제하에 정보의 손실을 최소화하면서 기계가 필요로 하는 모든 정보를 최대한 보존하는 방식으로 설계된다. ACoM 기술의 실현을 위해 기존의 오디오 코딩 기술을 활용할 때 고려해야 할 주요 과제는 다음과 같다.

첫 번째로, 압축된 오디오 데이터가 기계 학습 모델과 효과적으로 통합될 수 있어야 한다. 단순한 오디오 신호의 복원을 넘어서 기계 학습 모델에 직접 입력 가능한 형태의 데이터 표현이 필요하다. 예를 들어, 오디오 신호에서 중요한 특징(Feature)을 미리 압축하여 전송하면 기계 학습 모델이 신호를 복원하지 않고 바로 분석하여 보다 빠르고 효율적인 처리가 가능하다. 또한 기계를 위한 잠재적 주요 특징 신호가 무엇인지도 정의될 수 있어야 하며 이를 활용한 부호화 기술 개발이 필요하다. 둘째, 인간의 청각으로 인지할 수 없는 초음파나 저주파 대역 등 새로운 주파수 영역을 효과적으로 압축하고 표현할 수 있는 기술이 개발되어야 한다. 이는 종전의 오디오 코덱이 가청 주파수 대역에만 초점을 맞춘 것과는 달리 산업용 기계 모니터링, 초음파 탐지 등에서 기계 분석을 위해 필요한 비가청 영역의 정보를 효과적으로 압축하면서도 정보 손실을 최소화하는 방법을 ACoM에서는 마련해야 하기 때문이다. 마지막으로, 기존 코덱보다 더 높은 압축 효율성을 달성해야 한다. ACoM은 다양한 센서 데이터를 다자간 통신으로 공유하는 환경에서 사용될 가능성이 크기 때문에, 높은 압축 효율을 달성하지 못하면 실제 시스템에서의 응용 가능성이 제한될 수 있다. 따라서, ACoM의 성공적인 기술 개발 및 표준화를 위해서는 기존 인간 중심 오디오 부호화 기술의 한계를 인식하고, 하위 호환성 확보, 메타데이터 활용, 기계 학습과의 연계성

강화, 높은 압축 효율 달성 등 다양한 기술적 도전 과제를 해결하는 것이 필수적이다.

Ⅲ. MPEG ACoM 표준화 동향

본 장에서는 MPEG ACoM의 개념과 표준화 목표 및 범위를 살펴보고 주요 사용 사례 및 요구 사항을 정리한다. 또한 MPEG에서의 표준화 진행 현황을 정리한다.

1. MPEG ACoM 개념과 목표

ACoM은 기계 기반 응용을 위한 오디오 및 관련 데이터 부호화를 다루는 MPEG의 신규 표준화 아이템이다. ACoM의 목표는 오디오 신호 또는 그로부터 추출된 특징을 효율적으로 압축하여, 복원 후에도 기계 학습 알고리즘이 해당 데이터를 활용해 다양한 작업을 수행할 수 있는 공통 비트스트림 형식을 정의하는 것이다. 즉, ACoM에서 정의한 비트스트림을 복호화하면 기계가 이해하거나 추가 처리하기에 적합한 신호나 특성이 얻어지며, 필요에 따라 인간 청취용 신호로도 활용될 수 있다. 또한 표준 형식에는 메타데이터를 포함하여, 오디오 또는 다차원 센서 데이터의 캡처 방법, 환경 등을 기술함으로써 데이터에 대한 부가 정보를 제공하도록 한다[3,10].

ACoM 표준에서 주로 고려되어야 할 요소는 낮은 지연 시간, 고효율의 데이터 압축, 실시간 동작 지원 및 기계 최적화 등이다. 예를 들어 자율주행차나 실시간 모니터링 시스템에서는 오디오 이벤트를 빠르게 인지해야 하므로, 코딩 지연을 최소화하고 실시간 동작성을 확보하는 것이 중요하다. 또 방대한 센서 및 오디오 데이터 스트림을 네트워크로 전송하거나 저장하려면 고효율의 압축이 필수적이며, 기존 압축 기술 대비 더 높은 압축 효율을 제공해야 한다. 또한, 기계 학습 모델 동작을 위한 최적화가 필요한데 이는 압축 과정에서 기계를 위한 정보 손실을 최소화하여 분석 성능을 저하시키지 않도록 하는 것을 의미한다. 예를 들어, 기존의 지각 최적화 오디오 코덱들이 인간의 청각 특성에 맞춘 손실 압축을 수행한 반면, ACoM은 기계 학습 알고리즘의 인식 품질을 새로운 척도로 삼아 부호화 성능을 평가하는 지표가 고려되어야 한다.

ACoM 표준화는 두 단계(Phase)로 나누어 추진될 계획이다. Phase 1에서는 특정 응용에 종속되지 않는 범용 데이터 코덱으로서, 오디오 신호(1차원 혹은 다차원) 또는 다차원 스트림(예: 의료 데이터)을 거의 무손실(Near-lossless)로 압축하는 기술을 개발한다. 이를 통해 다양한 데이터를 정보 손실 없이 효율적으로 교환할 수 있는 표준 비트스트림 형식을 우선 확보하고자 한다. Phase 2에서는 특징 추출 기반 부호화(Feature-based coding) 방식을 도입한다. 즉, 특정 작업에 최적화된 특징들을 추출하여 손실 압축하는 방법을 포함시키며, 응용 분야별로 최적화된 코덱 모드를 정의할 수 있다. Phase 2의 결과물은 데이터 압축 효율을 극대화하면서도, 특정 기계 작업에 유용한 표현을 제공하는 것을 목표로 한다. 이러한 2단계 로드맵은 표준화를 단계적으로 진행함으로써 우선 산업계에 즉시 활용 가능한 형식을 제공하고(Phase 1), 향후 기술 발전에 따라 보다 진화된 기계 기반 응용을 위한 부호화(Phase 2) 방식을 도입하려는 목적이다.

2. ACoM의 범위와 시스템 개요

ACoM 표준의 범위에는 일반 오디오 신호뿐 아니라 다차원 센서 스트림 및 오디오 특징 데이터까지 포함된다. 예를 들어 마이크 배열로 수집한 공간 오디오, 의료용 센서 데이터(뇌파 신호 등)처럼 시간-채널에 따라 다차원으로 구성된 신호도 ACoM의 대상이며, Phase 2에서는 사람이 들을 수 없는 주파수 대역의 신호나 이미 추출된 특징 벡터까지도 입력으로 취급한다. 이러한 다양한 입력을 처리하기 위해 ACoM 시스템은 적응형 전처리/후처리, 특징 추출 모듈, 메타데이터 처리 모듈 등을 포함한 코덱 아키텍처를 갖춘다. 그림 1은 MPEG에서 논의 중인 ACoM 코덱의 개략적인 구조를 Phase 1(흰색 블록)과 Phase 2(회색 블록) 관점에서 도식화한 것이다[11]. 점선은 필요에 따라 선택적으로 연결될 수 있는 경로와 블록을 의미한다.

ACoM 부호화기는 입력에 따라 오디오 부호화(적응형 전처리, 부호화) 또는 특징 부호화(특징 추출, 변환, 부호화)의 두 경로로 동작하며, 부호화된 오디오/특징 데이터와 메타데이터를 비트스트림으로 출력한다. 복호화기에서도 대

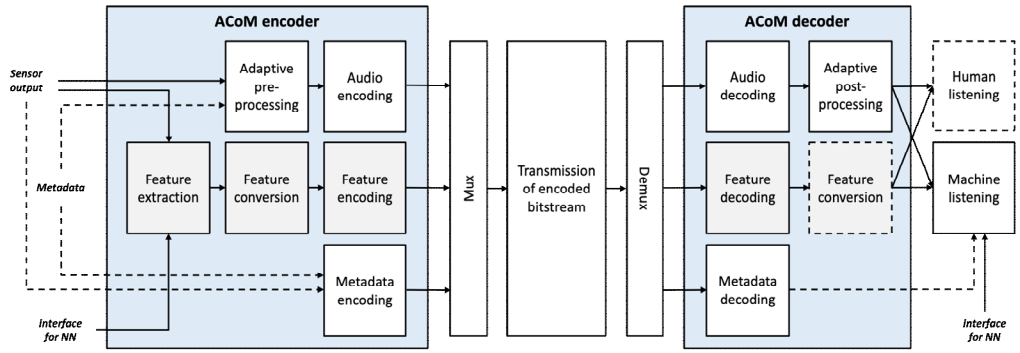

그림 1. MPEG ACoM 시스템 구조 예시[11]
Fig. 1. An example of potential ACoM architecture (adapted from[11])

응되는 경로를 통해 신호와 메타데이터를 복원하며, 복호화 후 필요시 적응형 후처리를 통해 인간 청취에 적합하거나 기계 분석에 최적화된 형태로 신호를 재구성한다.

ACoM 코덱은 오디오 신호 코딩 경로와 특징 코딩 경로를 모두 포괄하는 유연한 구조를 갖는다. 오디오 코덱 모드에서는 입력 오디오에 대해 필요한 경우 적응형 전처리를 수행하여 기계 분석에 적절한 형식이나 구성 요소로 신호를 변환한 뒤 압축하며, 디코딩 후에는 적응형 후처리를 통해 구성 요소들을 변환하여 목적에 맞게 활용 가능하도록 재구성한다. 적응형 전/후처리는 초음파 신호 등 다양한 신호에 대응하기 위해 주파수 대역 분할 및 시프트 등을 수행할 수 있으며, 기계 분석이나 인간 청취에 모두 적용될 수 있다. 반면 특징 코덱 모드에서는 별도의 특징 추출기(예: 신경망 기반)를 통해 입력 신호로부터 특징 벡터를 얻고, 이를 특징 변환 및 부호화 모듈을 통해 압축한다. 특징 코덱 모드는 Phase 2에 해당하며, 다양한 작업에 최적화된 특징을 전송하여 기계 학습 기반의 분석 성능을 극대화하는 것이 목적이다. 이처럼 ACoM 표준은 단순 신호 압축을 넘어서 기계 학습과 부호화의 접목을 추구하는 방향으로 발전하고 있다. 아울러 ACoM 비트스트림에는 앞서 언급한 메타데이터(예: 센서 위치, 환경 정보, 어노테이션 등)도 함께 부호화되어 전송되며, 복호화 측에서 기계 학습 모듈이 이를 활용하여 보다 정확한 분석을 수행할 수 있게 한다.

3. 사용 사례

MPEG ACoM에서는 표준화를 위해 다양한 사용 사례

표 1. MPEG ACoM의 사용 사례 및 적용 분야
Table 1. Use cases and applications of MPEG ACoM

Use Case		Application
UC1	Predictive Maintenance	Industrial
UC2	Process Control	
UC3	In-line Testing	
UC4	End-of-line Testing	
UC5	Traffic Monitoring and Control	Site Monitoring
UC6	Construction Site Monitoring	
UC7	Speech Recognition and Acoustic Scene Analysis	Life-logging / Contents Service
UC8	Timed Medical Data	Medical
UC9	Flexible Medical Data	
UC10	UGC Analysis	Contents Service
UC11	Live Stream Content Analysis	
UC12	Artistic creation	

(Use case)를 정의하여 적용 가능 분야를 구체화하였다[11]. 표 1은 현재까지 논의된 대표적인 ACoM 사용 사례 12가지를 분야별로 정리한 것이다. 산업, 도시 모니터링, 콘텐츠 서비스, 의료 등 여러 도메인에 걸쳐 기계 기반 오디오 데이터 처리 기술의 수요가 존재함을 알 수 있다.

먼저 산업 현장 분야의 UC1~UC4는 공장 기계의 예측 유지보수(Predictive Maintenance), 공정 제어(Process Control), 제품 테스트를 위한 오디오 모니터링 등을 다룬다. 예를 들어 여러 개의 센서로 기계 동작음을 수집하여, 이상 징후가 있는지 기계 학습으로 판정하는 시나리오이다. 이러한 경우 ACoM 코덱은 다수 채널의 센서 데이터를 무손실에 가깝게 압축하면서도, 필요한 경우 인간이 재생해서 들을 수 있도록 하는 기능이 요구된다. 또한 센서의 위치나 운영 조건, 타임스탬프 등의 메타데이터를 함께 저장하

여 필요한 정보를 제공하거나 예측 유지를 위한 분석 효율을 높일 수 있다.

다음으로 UC5 교통 모니터링(Traffic Monitoring and Control)과 UC6 건설 현장 모니터링(Construction Site Monitoring)은 스마트 시티/환경 모니터링에 해당한다. 도로 교통량이나 공사장 소음을 상시 모니터링하면서 교통 흐름 모니터링, 이상음 탐지, 소음 공해 관리 등을 수행하는 경우로, 대역폭이 제한된 무선 센서망을 통해 데이터를 전송해야 하므로 고압축 효율이 중요하며 저전력 환경에서의 실시간에 가까운 처리가 필요하다. ACoM은 이러한 환경음 데이터를 효율적으로 압축하고, 시간 동기화를 보장하여 여러 센서의 정보를 종합적으로 분석할 수 있게 하는 것을 목표로 한다.

UC7 음성 인식 및 음향 장면 분석(Speech Recognition and Acoustic Scene Analysis)은 다채널 마이크 신호를 활용하여 음성 인식, 감정 인식, 감정 분석, 화자 인식 및 환경 음향 분석과 같은 다양한 기계 학습 작업을 지원하는 사례이다. 이 사용 사례에서는 특히 언어, 화자 정보, 음성의 문자 전사본(Transcription)과 같은 메타데이터를 함께 인코딩할 수 있어야 하며, 환경 조건(예: 온도, 습도, 기압) 정보 또한 메타데이터로 제공될 수 있다. 또한 음향 장면 분석을 통해 주변 환경 소리의 특징까지 분석 가능한 형태로 제공되어야 한다. 이러한 분석 작업들은 기기의 자동 음성 인식 서비스나 감정 인식, 환경 소음 분석 등 다양한 응용 분야에서 활용될 수 있다. 따라서 ACoM 코덱은 오디오 신호뿐 아니라 다양한 메타데이터를 효율적으로 압축하고, 복호화 후에도 기계 분석 성능을 저하시키지 않아야 하며, 필요시 인간이 이해 가능한 형태(예: 음성 텍스트 전사본)로도 제공 가능해야 한다.

UC8, UC9는 의료 데이터(Medical Data) 분야로, 시간에 따른 생체음향 신호(예: 심장 박동음)나 의료 센서 스트림(EEG 등)을 압축 저장하는 사례이다. 의료 데이터는 손실 없는 보존이 매우 중요하지만, 장기간 기록 시 데이터량이 방대해지므로 효과적인 압축이 필요하다. ACoM은 무손실, 고정밀 압축을 통해 의료 신호의 중요한 세부 정보를 보존하면서 저장 효율을 높이는 솔루션을 제공할 수 있다.

마지막으로 콘텐츠 서비스 및 콘텐츠 생성 분야의 사용 사례로는 UC10 사용자 생성 콘텐츠 분석(User-Generated Content Analysis), UC11 실시간 스트리밍 콘텐츠 분석(Live Stream Content Analysis) 및 UC12 예술 창작(Artistic Creation)이 있다. UC10 및 UC11은 플랫폼에 업로드된 콘텐츠나 라이브 스트림 오디오에서 유용한 정보를 실시간으로 추출하여 사용자 경험을 향상시키거나 개인화 추천 서비스에 활용하는 사례이다. 특히 이 과정에서 오디오 임베딩(Audio Embedding)과 같은 기계 학습 기반의 특징 추출이 사용되며, 오디오의 유형(음악, 음성 등)과 같은 메타데이터를 포함하여 정확한 시점 정보와 같은 상세한 메타데이터도 제공할 수 있어야 한다. UC12는 음악 작곡, 음향 효과 생성, 잡음 제거, 음성 제거, 개별 음원 분리 및 편집과 같은 창작 콘텐츠 제작을 위한 응용을 포함한다. 이 사용 사례에서는 오디오뿐 아니라 자막, 대화, 음악 악보 등 콘텐츠 제작에 필수적인 메타데이터를 포함한 부호화를 지원해야 하는 기술이 요구된다. 이러한 콘텐츠 서비스 및 생성 분야의 응용 사례들에서도 하나의 ACoM 스트림에서 기계 분석을 위한 특징 데이터와 동시에 인간의 청취 또는 콘텐츠 소비를 위한 신호나 메타데이터를 효율적으로 지원할 수 있는 하이브리드 형태의 부호화가 필요하다.

이상과 같이 다양한 사례에서 공통적으로 등장하는 요구는 다채널 오디오의 효율적 압축, 필요한 경우 무손실 수준의 정확도, 메타데이터 부가 정보 전송, 기계 분석에 유용한 특징 보존 등을 들 수 있다. ACoM 표준은 이러한 요구 사항을 만족하는 표준을 제공함으로써, 각 산업 분야에서 상호운용성과 효율성을 크게 높일 것으로 기대된다.

4. ACoM 표준 요구 사항

MPEG에서는 ACoM의 기술적 방향을 구체화하기 위해 다양한 요구 사항(Requirements)을 정의하였다[11]. 요구 사항들은 앞서 살펴본 사용 사례 전반에 걸쳐 공통적으로 고려되어야 할 핵심 요소들을 반영한다. 주요 ACoM 표준 요구 사항은 다음과 같다.

1) 압축 효율성: ACoM으로 생성된 압축 비트스트림의 크기는 반드시 기존의 전통적 오디오 무손실 코딩보다도 작아야 한다. 이는 곧 ACoM이 새로운 부호화

기법을 통해 데이터 내 중복성을 최대한 제거하고, 인간 청각이 고려되지 않는 만큼 불필요한 정보까지 효과적으로 압축함을 의미한다.

2) 단일/다중 작업 지원: ACoM 부호화 기술은 하나의 비트스트림으로 단일 작업뿐 아니라 복수의 분석 작업을 지원할 수 있어야 한다. 예를 들어 하나의 압축 특징 스트림으로부터 오디오 이벤트 검출과 객체 수준 분류, 행동 인식 등 여러 알고리즘이 각자 필요한 정보를 얻어낼 수 있어야 하며, 만약 특정 한 가지 작업만 대상일 때는 그에 맞게 최적화(더 높은 압축 또는 낮은 복잡도)할 수 있는 유연성을 가져야 한다.
3) 작업별 성능 가변성: 동시에 여러 작업에 데이터를 제공하는 경우, 일부 중요 작업의 성능을 우선 보장하는 기능이 필요하다. 예컨대 긴급성이나 중요도가 높은 분석은 낮은 지연과 높은 정확도를 확보하고, 부가적인 작업에 대한 성능은 다소 저하되는 방식으로 우선 순위에 따른 인코딩 품질 조절을 지원해야 한다. 이를 위해 비트스트림 내에 작업별 우선도 정보를 표시하는 플래그 등이 고려될 수 있다.
4) 무손실 및 유연한 손실 압축: Phase 1의 목표인 무손실 부호화는 기본 요구 사항이며, 다채널 간 상관관계까지 활용해 압축 효율을 높여야 한다. 아울러, 각종 메타데이터는 거의 무손실로 압축하여 측정 정밀도 이내의 정확도를 유지해야 한다.
5) 프라이버시 보호: 특정 응용(특히 UC7 등)에서는 ACoM 비트스트림으로부터 원 음성이나 화자 정보가 재구성이 불가 하도록 하는 것이 요구된다. 이는 개인정보를 보호하기 위한 것으로, 예를 들어 음성 인식용 특징만 보내고 음색 등 화자와 관련된 정보는 버리거나 암호화하는 등의 구현 옵션을 고려할 수 있다.
6) 메타데이터 인코딩: 녹음 설정이나 환경에 대한 메타데이터를 비트스트림에 함께 담아 전송할 수 있어야 한다. 예를 들어 센서의 위치나 종류, 녹음 당시의 조건, 또는 UC7의 경우 해당 음성의 전사 주석(Annotations of Text Spoken) 등이 이에 해당한다. UC9~UC12의 경우에도 콘텐츠에 대한 오디오 레이블 등의 메타 정보가 유용할 수 있으며, ACoM 형식은 이를 포함할 수 있어야 한다.
7) 인간-기계 혼합 소비: ACoM에서는 하나의 공통 비트스트림을 기계와 인간이 모두 활용할 수 있어야 한다. 이를 위해 필요한 경우 동일 비트스트림으로부터 기계용 특징과 인간용 신호를 모두 복원할 수 있는 하이브리드 부호화 방식을 요구하며, 일부 사용 사례(UC1, UC5, UC6, UC7, UC10, UC11)에서는 이를 나타내기 위해 플래그 정보가 사용될 수 있다.
8) 엣지 컴퓨팅 및 동기화: ACoM에서는 여러 센서로부터 얻어진 데이터를 처리하고 동기화하는 시나리오를 지원하기 위해 엣지 컴퓨팅(Edge Computing) 및 동기화를 요구 사항으로 제시하고 있다. 이를 위해 분산된 센서 노드에서 수집된 오디오 및 다차원 데이터를 엣지 컴퓨팅 및 네트워크를 통해 효율적으로 전송하고, 획득된 데이터가 중앙 시스템에서 정렬되어 통합 분석될 수 있도록 타임스탬프 기반의 신호 동기화 기능이 요구된다.
9) 비가청 주파수 대응: 인간에게는 들리지 않지만 기계 분석에 유용한 초음파 등 비가청 신호도 처리 대상으로 포함될 수 있다. 따라서 ACoM 코덱은 기존 오디오 코덱이 다루지 않는 주파수 대역까지도 포괄하거나, 별도의 초음파 채널도 압축하는 유연성을 고려해야 한다.
10) 채널 확장성: ACoM은 적용 분야에 따라 입력 채널의 개수가 한 개에서 수십 개까지 다양할 수 있으므로 채널 수에 대한 확장성을 가져야 하며, 각 사용 사례별로 필요한 최소/최대 채널 수를 정하고 그 범위 안에서 효율적인 압축을 제공하도록 설계되어야 한다.

이러한 요구 사항들은 ACoM 표준화의 기술적 가이드라인으로 작용하여, 제안되는 코덱이 어떤 기능과 성능을 갖추어야 하는지를 정의한다. 현재 MPEG에서는 이러한 요구 사항들을 바탕으로 ACoM 코덱 기술에 대한 품질 평가 방법, 객관적 지표 설정, 데이터 세트, 메타데이터 형식 등 세부적인 항목들에 대한 구체화가 진행되고 있다.

5. MPEG 표준화 진행 현황 및 계획

MPEG에서 ACoM 표준화는 2022년 10월 MPEG 140차 회의에서 차세대 오디오 코딩 주제로 Fraunhofer IDMT에 의해 처음 제안되었으며, MPEG 오디오 분과(WG6)와 기술 요구 분과(WG2)에서 공동으로 사용 사례 및 요구 사항을 논의하는 작업이 시작되었다[12,13]. 이후 2024년 4월 MPEG 146차 회의에서 해당 논의가 WG6으로 이관되어 본격적인 표준화 기술 탐색(Exploration) 단계에 돌입하였으며[14], 2024년 7월 147차 회의에서는 CfE(Call for Evidence) 단계로 진입하기로 결정되어 CfE를 위한 준비 작업 계획이 구체화 되었다[15]. 147차 회의에서는 콘텐츠 서비스와 의료 모니터링 관련 기고서가 제출되어 사용 사례에 반영되었으며[16-18], 2024년 11월 148차 회의에서도 기타 ACoM 서비스를 위한 적응형 전/후처리 기술 등 다양한 제안들이 제출 되었다[19,20]. 2025년 1월 MPEG 149차 회의에서는 ACoM CfP(Call for Proposals) 작성이 주요 안건으로 다루어져, CfP 초안 및 테스트 계획, 일정 등이 논의되었다[21]. 특히 메타데이터 입력 포맷, 데이터 세트, 성능 평가 지표 등에 관한 세부 논의가 진행 중이며, 2025년 4월 150차 회의를 통해 CfP 안을 작성하고 정식 CfP를 발행할 계획을 수립 중이다. CfP 발간 후에는 이에 대한 응답을 받고 평가할 예정이다. 표준 후보 기술들에 대한 평가와 병합 작업(Core Experiment)을 수행한 뒤, 표준 초안을 작성하게 되며, 이후 MPEG의 일반적인 표준화 절차에 따라 위원회 초안(Committee Draft, CD), 표준 초안(Draft International Standard, DIS), 최종안(Final Draft International Standard, FDIS) 승인의 과정을 거쳐 국제 표준으로 공식 채택될 예정이다.

Ⅳ. 토의 및 결론

본 논문에서는 MPEG ACoM의 표준화 동향과 기술적 요구 사항을 분석하고, 기존 오디오 코딩 기술의 한계를 극복하기 위한 방향성을 논의하였다. 기존 오디오 코딩 기술은 인간의 청취 경험을 최적화하는 방식으로 발전해 왔으나, AI 및 기계 학습의 발전과 함께 기계가 이해하고 분석할 수 있는 오디오 부호화 방식이 요구되고 있다. ACoM은 이러한 필요성을 반영하여 기계 중심의 오디오 데이터 압축 및 처리 방식을 정의하고자 한다. 이에 따라 본 논문에서는 기존 오디오 코딩 기술의 한계를 분석하고, ACoM이 해결해야 할 주요 기술적 과제를 도출하였다. 기존 오디오 코딩 기술은 인간 청각을 고려한 손실 압축 방식이므로, 기계 학습 및 분석에 필요한 비가청 신호나 미세한 음향 패턴을 유지하는 데 한계가 있다. 기계 학습 기반 오디오 분석을 위해서는 초다채널 및 초광대역 오디오 데이터를 처리할 수 있는 새로운 접근 방식이 필요하며, ACoM은 기계 학습 및 분석을 위한 최적화된 오디오 부호화 기술을 통해 이러한 문제를 해결하고자 한다. 현재 MPEG에서는 ACoM 표준화를 위한 기술 요구 사항을 정의하고 있으며, 산업용 기계 모니터링, 음성 인식, 의료 데이터 분석, 콘텐츠 분석 등 다양한 분야에서 활용 가능성을 고려하고 있다. 표준화가 완료되면 기계 간의 오디오 데이터 교환 및 처리 효율성이 증가하고, AI 기반 오디오 분석 기술의 발전이 가속화될 것으로 기대된다.

ACoM의 성공적인 표준화를 위해서는 여러 기술적 과제를 해결해야 한다. 우선, 기계 학습 및 AI 분석을 위해서는 압축된 데이터에서도 분석 성능이 유지될 수 있도록 정보 손실을 최소화하는 기술이 필요하다. 이를 위해 기존의 무손실 및 손실 압축 방식과 차별화된 AI 기반 데이터 압축 및 인코딩 기법이 요구되며, 메타데이터를 적극적으로 활용해 기계 학습 모델과의 연계성을 높이는 방안이 마련되어야 한다. 또한, 스마트 팩토리, 자율주행, 의료 모니터링 등과 같이 실시간성이 중요한 산업 분야에서는 엣지 디바이스에서도 원활하게 동작할 수 있도록 부호화 기술의 경량화와 최적화가 필수적이다. 아울러, 기존 오디오 코딩 기술과의 호환성을 유지하면서도 기계 중심 분석을 위한 새로운 압축 방식과 데이터 구조를 포함하는 방향으로 기술이 발전해야 한다.

향후 ACoM 표준화가 완료되면 기계 학습 응용 분야에서 데이터 형식의 통일을 통해 산업 간 상호운용성을 높일 수 있으며, 향상된 압축 효율은 저장 및 전송 비용 절감뿐 아니라 분석 시스템의 실시간성을 향상시키는 데 기여할 수 있다. 더 나아가, 이러한 기술은 기기나 클라우드 기반의 지능형 인식 서비스를 가능하게 하여 새로운 비즈니스 모

델 창출로 이어질 것으로 기대된다. 결론적으로 MPEG의 ACoM 표준은 아직 해결해야 할 기술적 과제가 남아 있지만, 다양한 산업 분야에서 효율적인 기계 지능 기반 데이터 활용을 가능하게 하는 기반 기술로 자리매김할 것이며, 인공지능 시대의 오디오 데이터 처리 요구에 부응하는 핵심 표준으로 발전할 것으로 전망된다.

참 고 문 헌 (References)

[1] Ericsson mobility report, Nov. 2023. [online] Available: https://www.ericsson.com/en/reports-and-papers/mobility-report/reports

[2] M. Kim, J. Skoglund, “Neural Speech and Audio Coding: Modern AI technology meets traditional codecs,” IEEE Signal Processing Magazine, 2025.
doi: https://doi.org/10.1109/MSP.2024.3444318

[3] J. Kang, S. Lim, S. Bae, S. Jeong, I. Jang, “Trends in the Standardization of AI-Based Media Coding Technology,” Electronics and Telecommunications Trends, 2025.
doi: https://doi.org/10.22648/ETRI.2025.J.400203

[4] ISO/IEC 13818-7:2006, “Information technology - Generic coding of moving pictures and associated audio information - Part 7: Advanced Audio Coding (AAC),” 2006.

[5] J. Valin, K. Vos, and T. Terriberry, “Definition of the Opus Audio Codec,” IETF RFC 6716, Sep. 2012.
doi: https://doi.org/10.17487/RFC6716

[6] ISO/IEC 23003-3:2012, “Information technology - MPEG audio technologies - Part 3: Unified speech and audio coding,” 2012.

[7] ISO/IEC 23008-3:2019, “Information technology - High efficiency coding and media delivery in heterogeneous environments - Part 3: 3D audio,” 2019.

[8] M. R. Schroeder and B. S. Atal, “Code-excited linear prediction (CELP): High-quality speech at very low bit rates,” in Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1985.
doi: https://doi.org/10.1109/ICASSP.1985.1168147

[9] M. Neuendorf et al., “MPEG Unified Speech and Audio Coding - the ISO/MPEG standard for high-efficiency audio coding of all content types,” in Proceedings of the 132nd Audio Engineering Society Convention, 2012.

[10] S. Byun, J. Seo, “Analysing Trends in Audio Coding for Machines,” Summer Conference, KIBME, 2024.

[11] Use Cases and Requirements on Audio Coding for Machines, N308, ISO/IEC JTC1 SC29/WG6, Jan. 2025.

[12] T. Sporer, “Proposal for New Work Item: Audio Coding for Machines,” m61162, Oct. 2022.

[13] Use Cases and Requirements for Audio Coding for Machines I(ACoM), N252, ISO/IEC JTC1 SC29/WG2, Oct. 2022.

[14] Use Cases and Requirements on Audio Coding for Machines, N252, ISO/IEC JTC1 SC29/WG6, April 2024.

[15] Workplan on Audio Coding for Machines, N269, ISO/IEC JTC1 SC29/WG6, July 2024.

[16] Z. Wang, B. Wang, M.-L. Champel, “Market and practical considerations (ACoM: AIGC Creation and Optimization),” m68401, ISO/IEC JTC1 SC29/WG6, July 2024.

[17] Z. Wang, B. Wang, M.-L. Champel, “Market and practical considerations (ACoM: Nursing-Baby Monitoring),” m68402, ISO/IEC JTC1 SC29/WG6, July 2024.

[18] Use Cases and Requirements on Audio Coding for Machines, N268, ISO/IEC JTC1 SC29/WG6, July 2024.

[19] Z. Wang, B. Wang, M.-L. Champel, “Some thoughts of ACoM architecture,” m69849, ISO/IEC JTC1 SC29/WG6, Nov. 2024.

[20] Y. Zhu, C. Huang, “Thoughts on ACoM used in training and inference stage,” m69915, ISO/IEC JTC1 SC29/WG6, Nov. 2024.

[21] Workplan on Audio Coding for Machines, N309, ISO/IEC JTC1 SC29/WG6, Jan. 2025.

저 자 소 개

임 우 택

- 2010년 : 광운대학교 전자공학과 공학사
- 2012년 : 광운대학교 전자공학과 공학석사
- 2025년 : 한국과학기술원 문화기술대학원 공학박사
- 2012년 ~ 현재 : 한국전자통신연구원 미디어부호화연구실 선임연구원
- ORCID : https://orcid.org/0009-0006-4640-4301
- 주관심분야 : 오디오 신호처리, 기계 학습

저 자 소 개

장 인 선

- 2001년 2월 : 충북대학교 전기전자공학부 학사
- 2004년 2월 : 포항공과대학교 컴퓨터공학과 석사
- 2018년 : 충남대학교 전자전파정보통신공학과 공학박사
- 2023년 ~ 2024년 : 미국 Indiana University, 방문연구원
- 2004년 8월 ~ 현재 : 한국전자통신연구원 미디어부호화연구실 책임연구원
- 주관심분야 : 오디오 부호화 및 신호처리, 기계 학습
- ORCID : https://orcid.org/0000-0003-2237-2668

백 승 권

- 2005년 : 한국과학기술원 정보통신공학부 공학박사
- 2005년 ~ 현재 : 한국전자통신연구원 미디어부호화연구실 책임연구원
- ORCID : https://orcid.org/0000-0002-6254-2062
- 주관심분야 : 오디오 신호처리, 음성/오디오 코덱

강 정 원

- 2003년 8월 : 조지아공과대학교 전자 및 컴퓨터공학과 공학박사
- 2003년 10월 ~ 현재 : 한국전자통신연구원 미디어부호화연구실 책임연구원
- ORCID : https://orcid.org/0000-0003-4003-4638
- 주관심분야 : 비디오 부호화, 오디오 부호화, 멀티미디어 처리