생성형 AI가 여는 차세대 미디어 기술의 지평 : 2026년도 한국방송·미디어공학회 하계교육 후기

강제원 조직위원장 (이화여자대학교)

요약

한국방송·미디어공학회는 2026년 8월 25일 “생성형 AI 원천 기술과 차세대 미디어 핵심 응용: 음성·영상·3D 콘텐츠 생성, 전달에서 신뢰성까지”를 주제로 2026년도 하계교육을 온라인으로 개최하였다. 이번 교육은 생성형 AI가 미디어 산업과 콘텐츠 기술 전반에 가져오는 변화를 이해하고, 그 기반이 되는 원천 기술과 응용 사례를 체계적으로 학습하도록 구성되었다. 오전에는 옴니모달 AI의 효율화 기술과 Diffusion 모델의 수학적 기반 및 영상 생성 응용을, 오후에는 음성 중심 상호작용 AI, 음성 생성과 딥보이스 탐지, 피드포워드 3D Gaussian Splatting, 그리고 3D Gaussian Splatting 압축 기술을 다루었다. 여섯 강연은 각각 독립된 주제를 다루면서도 ‘생성 → 전달 → 신뢰성’이라는 미디어 처리의 전 과정을 하나의 흐름으로 엮었다. 사전등록 기준 대학·연구소·방송사 등 21개 기관에서 58명이 등록하였으며, 이 가운데 약 3분의 2가 대학원생이었다. 본 교육은 생성형 AI가 더 이상 개별 연구 주제가 아니라 방송·미디어 기술의 공통 기반이 되었음을 확인하는 자리였다.

행사 개요

2026년도 한국방송·미디어공학회 하계교육은 2026년 8월 25일(화) 오전 10시부터 오후 5시 30분까지 줌(Zoom)을 활용한 온라인 방식으로 진행되었다. 교육의 주제는 "생성형 AI 원천 기술과 차세대 미디어 핵심 응용: 음성·영상·3D 콘텐츠 생성, 전달에서 신뢰성까지"였다. 프로그램은 강제원 교수(이화여자대학교)가 조직위원장을, 김선옥 교수(한국항공대학교)가 프로그램위원장을 맡고, 진경환(고려대학교), 오지형(중앙대학교), 조성인(서강대학교), 최진우(경희대학교), 홍성은(성균관대학교), 엄찬호(중앙대학교), 고현석(한양대학교), 최종원(중앙대학교) 교수가 프로그램위원으로 참여하여 기획하였다.

교육 프로그램은 생성형 AI의 기반 기술에서 출발하여 음성, 영상, 3D 콘텐츠로 응용 범위를 넓혀 가도록 배치되었다. 오전 세션에서는 옴니모달 파운데이션 모델의 효율화와 Diffusion 모델의 원리를 다루어 이후 강연의 이론적 토대를 마련하였다. 오후 첫 세션은 음성과 오디오 생성을 중심으로 인간과 AI의 상호작용 및 딥보이스의 사회적 위협을 다뤘고, 마지막 세션은 3D Gaussian Splatting(3DGS)의 생성과 압축을 다루었다.

강연 내용

Foundations of Efficient Omnimodal AI
/ 홍성은 교수 (성균관대학교)

첫 강연에서 홍성은 교수(성균관대학교)는 텍스트를 넘어 이미지, 비디오, 오디오를 함께 이해하는 옴니모달 AI의 효율화 기술을 소개하였다. 모델 규모와 입력 컨텍스트가 커질수록 연산량, 메모리, 에너지 비용이 급격히 늘어나 실제 서비스와 온디바이스 환경에서의 활용에 제약이 따른다는 점이 출발점이었다. 강연자는 효율화 기술을 컨텍스트 압축, 모델 압축, 캐시·메모리 효율화, 적응적 연산, 시스템·서빙 최적화의 다섯 축으로 정리하여 체계적인 조망을 제시하였다.

구체적인 사례로는 비전-언어 모델의 시각 토큰 중복성 문제를 다룬 ZOO-Prune이 소개되었다. 이 방법은 추가 학습 없이 0차(zeroth-order) 기울기 추정으로 토큰이 출력에 미치는 민감도를 구하고, 이를 다양성 기준과 결합하여 중요한 토큰만 남긴다. 발표에 따르면 최대 94.4%의 시각 토큰을 제거하면서 최대 2.3배의 종단 간 추론 가속을 얻었다. 강연 말미에는 여러 모달리티 간 중복을 줄이는 옴니모달 압축과, 미래 프레임 없이 실시간으로 추론해야 하는 스트리밍 비디오 이해가 다음 연구 방향으로 제시되었다. 라이브 방송과 스마트 기기에서 대형 모델을 운용해야 하는 미디어 분야에 효율화가 선택이 아닌 전제 조건임을 보여주는 강연이었다.

Diffusion models for image and video generation
차은주 교수 (숙명여자대학교)

차은주 교수(숙명여자대학교)는 생성형 AI의 핵심 기술로 자리 잡은 Diffusion 모델의 수학적 기반을 심도 있게 다루었다. 강연은 에너지 기반 모델, VAE, GAN, Normalizing Flow로 이어지는 생성 모델의 발전사를 개관한 뒤, 데이터에 점진적으로 잡음을 더하는 순방향 과정과 잡음으로부터 데이터를 복원하는 역방향 과정을 설명하였다. 특히 역방향 과정을 변분(variational), 스코어 기반(score-based), 플로우 기반(flow-based)의 세 관점에서 비교하며, 세 관점이 결국 하나의 모델을 서로 다른 방식으로 기술한 것이라는 점을 보여주었다. DDPM에서 확률미분방정식(SDE) 기반 스코어 모델, 플로우 매칭으로 이어지는 설명은 수식과 그림을 함께 제시하여 학생 참가자도 이해하기 쉽도록 구성되었다.

후반부에서는 이론이 제어 가능한 생성으로 연결되는 과정을 다뤘다. 입력 영상의 정체성을 보존하면서 텍스트 지시에 따라 내용을 바꾸는 텍스트 기반 영상 편집의 과제가 소개되었고, 시공간적 일관성이 요구되는 비디오 생성으로의 확장이 논의되었다. 강연자는 참조 비디오와 피사체 이미지를 결합하여 학습 없이 개인화 비디오를 생성하는 연구(SA-V2V)를 사례로 들며, 신뢰할 수 있는 스코어 추정이 제어 가능한 생성의 핵심이라는 점을 강조하였다.

Toward Natural Human-AI Interaction: Spoken, Multimodal, and Conversational AI
박세진 교수 (경희대학교)

오후 첫 강연에서 박세진 교수(경희대학교)는 음성·영상·텍스트를 함께 활용하는 멀티모달 인간-AI 상호작용을 다루었다. 강연자는 음성 언어 모델(Spoken Language Model)이 음성 단위 기반 모델, 텍스트-음성 결합 모델을 거쳐 실시간 종단 간 음성 비서로 발전해 온 과정을 정리하고, 긴 호흡의 음성을 생성하는 장문 음성 생성 연구를 소개하였다. 이어 사용자의 얼굴과 음성을 함께 입력받아 얼굴과 음성으로 응답하는 대면 대화 모델, 입 모양까지 함께 번역하는 시청각 음성 번역(AV2AV) 등 시청각 상호작용 연구가 소개되었다.

특히 대화형 AI가 '무엇을 말할지'에만 최적화되어 있고, '어떻게', '언제' 말할지는 상대적으로 덜 다루어졌다는 지적이 인상적이었다. 강연자는 감정 선호 최적화를 활용한 공감형 대화 연구를 소개하고, 운율과 어조, 제스처와 시선 같은 비언어적 신호까지 아우르는 통합적 상호작용을 향후 과제로 제시하였다. 또한 상호작용의 변화를 예측하는 '상호작용 월드 모델'을 미래 방향으로 언급하였는데, 이는 디지털 휴먼과 AI 진행자 등 방송 콘텐츠에도 직접 연결되는 주제이다.

Advanced Speech Generation and Deep Voice Detection
김지훈 교수 (중앙대학교)

김지훈 교수(중앙대학교)는 음성 생성 기술의 발전과 그 이면인 딥보이스 문제를 함께 다뤘다. 강연은 내비게이션, 유튜브 콘텐츠, 음성 비서 등 일상 속 음성합성(TTS) 사례에서 출발하여, 텍스트 분석·음향

모델·보코더로 이어지는 신경망 TTS의 구조와 종단 간 TTS로의 발전을 차례로 설명하였다. 이어 화자의 목소리를 다른 화자의 것으로 바꾸는 음성변환(Voice Conversion) 기술을 발화 내용과 화자 정보를 분리하는 문제를 중심으로 소개하였다.

후반부는 이러한 기술이 악용된 오디오 딥페이크의 탐지에 할애되었다. 강연자는 생성된 오디오에 남는 고유한 흔적을 찾는 초기 방법부터 종단 간 모델, 자기지도 음성 모델 기반 탐지까지를 정리하였다. 또한 실험실 데이터에서는 성능이 높지만 실제 환경 데이터에서는 성능이 크게 떨어지는 일반화 격차를 주요 한계로 지적하고, 조작된 구간을 찾아내는 구간 탐지와 판단 근거를 제시하는 해석 가능한 탐지를 최근 동향으로 소개하였다. 보이스피싱 피해 보도 사례를 함께 다루며 기술 개발의 양면성과 책임 있는 활용을 강조한 점은, 교육 주제의 '신뢰성'을 가장 직접적으로 보여준 대목이었다.

Scalable Feed-Forward 3D Gaussian Splatting

박은병 교수 (연세대학교)

박은병 교수(연세대학교)는 "신경망이 최적화기를 이길 것"이라는 문제의식으로 피드포워드 3DGS를 다루었다. 3DGS는 빠른 학습과 실시간 고품질 렌더링으로 주목받아 왔으나, 장면마다 개별 최적화를 거쳐야 하고 데이터 기반 사전지식을 활용하지 못한다는 한계가 있다. 강연자는 다시점 영상에서 한 번의 추론으로 3D 장면을 생성하는 pixelSplat, MVSplat, GS-LRM 등의 흐름을 소개하고, 입력 시점과 해상도가 늘어날수록 트랜스포머의 연산량이 급격히 커지는 확장성 문제를 짚었다.

이를 해결하기 위한 방향으로 영상과 3D 표현을 분리하는 설계, 반복적 개선을 수행하는 iLRM, 인접 시점만 참조하는 지역 어텐션, 다중 해상도 피라미드 트랜스포머 등이 소개되었다. 카메라 자세 정보 없이 동작하는 포즈 프리(pose-free) 방식과, 픽셀마다 가우시안을 배치하는 대신 적은 수의 토큰으로 장면을 표현하는 비픽셀정렬 방식도 다루어졌다. 픽셀 정렬 방식에서는 성능 저하 없이 90% 이상의 가우시안을 제거할 수 있다는 결과는, 표현의 효율성이 다음 강연의 주제인 압축과 직접 연결됨을 보여주었다.

Compression in 3D Gaussian Splatting: A Survey of Methods, Trends, and Future Directions

배성호 교수 (경희대학교)

마지막 강연에서 배성호 교수(경희대학교)는 3DGS의 저장 용량과 메모리 문제를 해결하기 위한 압축 기술을 종합적으로 정리하였다. 3DGS는 수백만 개의 가우시안 입자로 장면을 표현하기 때문에 데이터 양이 크고, 이는 실감형 미디어의 저장과 전송에서 큰 제약이 된다. 강연자는 기존 연구를 가우시안 간 관계를 고려하지 않는 비구조적 압축과, 앵커·그리드 등 공간적 구조를 활용하는 구조적 압축으로 나누고, 가지치기(pruning), 양자화, 엔트로피 부호화 등 세부 기법을 분류하였다.

또한 대표 데이터셋에서의 성능 비교를 통해 압축률과 렌더링 품질 사이에 뚜렷한 상충 관계가 있으며, 앵커 기반의 구조적 방법이 유망한 성능을 보인다는 점을 제시하였다. 다만 압축 성능이 데이터셋 특성에 크게 좌우된다는 점도 함께 지적되었다. 표준화된 평가 체계와 동적 장면 압축 등 앞으로의 과제는

MPEG 등에서 진행중인 3D 미디어 표준화와도 맞닿아 있어, 방송·미디어 분야 연구자들에게 특히 시사하는 바가 컸다.

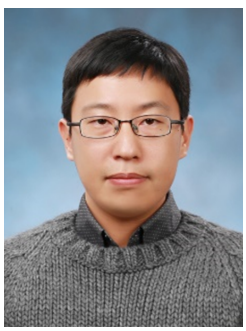
행사 후기

이번 하계교육은 생성형 AI를 단일 기술이 아닌 미디어 처리의 전 과정에 걸친 기반 기술로 바라보았다는 점에서 의의가 크다. 옴니모달 모델의 효율화와 Diffusion 모델의 원리에서 출발하여 음성과 3D 콘텐츠의 생성으로, 다시 압축과 전달, 딥보이스 탐지를 통한 신뢰성으로 이어지는 구성은 교육 부제가 내건 '생성, 전달에서 신뢰성까지'를 충실히 구현하였다. 각 강연이 기초 원리와 강연자의 최신 연구 성과를 함께 다루어, 처음 입문하는 대학원생과 현장의 연구자 모두에게 유익한 자리가 되었다.

여섯 강연에서 공통적으로 드러난 흐름도 있었다. 첫째, 모델의 성능보다 효율과 확장성이 주요 화두로 부상하였다. 시각 토큰 가지치기, 피드포워드 3DGS의 토큰 설계, 3DGS 압축은 모두 '얼마나 적은 자원으로 같은 품질을 얻을 것인가'라는 같은 질문에 답하고 있었다. 둘째, 별도의 추가 학습 없이 기존 모델을 활용하는 학습불필요(training-free) 접근이 여러 강연에서 등장하였다. 셋째, 생성 품질이 높아질수록 그 결과를 검증하고 악용을 막는 기술의 중요성도 커지고 있다는 점이 확인되었다.

방송·미디어 분야의 관점에서 보면, 이번 교육에서 다룬 기술은 현장 적용을 앞두고 있다. 실시간 스트리밍 이해와 음성 중심 대화 AI는 생방송과 디지털 휴먼에, Diffusion 기반 영상 생성은 콘텐츠 제작에, 3DGS의 생성과 압축은 실감형 미디어 서비스에 연결된다. 딥보이스 탐지와 같은 신뢰성 기술은 공공성과 신뢰를 중시하는 방송 매체가 더욱 주목해야 할 영역이다. 다만 실험실 성능과 실제 환경 성능 사이의 격차, 데이터셋에 따라 달라지는 압축 성능 등은 현장 적용을 위해 학계와 산업계가 함께 풀어야 할 과제로 남아 있다.

저 자 소 개



강 제 원

- 2012년 : University of Southern California 전자공학과 박사
- 2014년 : Qualcomm Multimedia RnD and Standard Team
- 2014년 ~ 2026년 : 이화여자대학교 전자전기공학과 교수
- 주관심분야 : 영상처리 및 압축, 멀티모달 AI, 미디어 AI