#초정밀 음성 인식과 언어의 경계 확장 #오디오 데이터의 지능형 분석 # 보이스 클로닝

5화

AI의 입, 소리를 지배하다
지능형 음성인식, 분석, 생성

글. 강자원 컴퓨터시스템응용기술사, KBS MNC(Media Network Center)팀

지난 호에서 다룬 AI의 '눈'이 시각 정보를 해독했다면, 이제 우리는 미디어를 완성하는 또 다른 축인 '소리'에 주목해야 한다. 단순한 파형(Waveform)에 머물렀던 오디오는 AI의 '입과 귀'를 통해 정교한 데이터 자산으로 거듭나고 있다. 이번 호에서는 음성인식(STT)과 생성(TTS) 기술이 방송 현장의 장벽을 허물고 제작진에게 '시간의 선물'을 안겨주는 공학적 여정을 살펴본다.

Part 1 서론: '보는 것' 그 이상의 가치, 오디오 데이터의 부활

컴퓨터 비전(CV) 기술이 영상 속 객체와 장면을 식별하며 AI의 '눈' 역할을 수행해 왔다면, 이제 우리는 미디어의 맥락을 완성하는 핵심 감각인 청각, 즉 AI의 '귀'와 '입'에 주목해야 한다. 시각 정보가 사건의 현상을 전달하는 일차적인 수단이라면, 소리는 그 사건의 이면에 숨겨진 의도와 감정, 그리고 눈에 보이지 않는 공간적 맥락을 채우는 결정적인 요소다. 하지만 오랜 시간 방송 제작 환경에서 오디오는 영상에 종속된 부수적인 신호로 취급받거나, 송출과 동시에 사라지는 휘발성 데이터에 머물러 있었던 것이 사실이다.

이제 지능형 음성 기술의 비약적인 발전은 이러한 패러다임을 완전히 뒤바꾸고 있다. 인공지능은 단순한 파형(Waveform)의 집합체였던 오디오 신호를 정교한 텍스트와 의미론적 구조를 가진 '비정형 데이터의 보고'로 재정의한다. 대화 속의 미세한 떨림과 톤의 변화에서 화자의 심리 상태를 읽어내고, 수만 시간의 아카이브 속에서 특정 키워드나 화자의 목소리를 단 몇 초 만에 찾아내며, 배경에 깔린 미세한 엠비언스(Ambiance)를 분석해 촬영 현장의 상황을 데이터화한다. 이는 소리가 단순한 기록물을 넘어, 영상 데이터와 대등한 가치를 지닌 독립적인 '디지털 자산'으로 격상되었음을 의미한다.

이러한 변화는 방송 엔지니어에게 기술적 숙련도를 넘어선 새로운 차원의 임무를 부여한다. 과거 엔지니어의 역할이 잡음 없는 깨끗한 소리를 수음하고 표준 규격에 맞춰 무결하게 송출하는 '신호 관리자(Signal Manager)'에 집중되었다면, AI 네이티브 시대의 엔지니어는 오디오 데이터에서 유의미한 메타데이터를 추출하고 이를 제작 시스템 전반에 흐르게 만드는 '미디어 언어 설계자(Media Language Architect)'로 진화해야 한다.
음성인식(STT) 기술로 추출한 텍스트를 검색 가능한 자산으로 전환하고, 이를 CMS(Contents Management System)나 MAM과 유기적으로 연동하여 제작 워크플로우를 혁신하는 과정은 엔지니어가 설계해야 할 새로운 공학적 영토다. 소리를 지배하는 자가 콘텐츠의 맥락을 지배한다는 확신을 바탕으로, 이제 본격적으로 지능형 음성 기술이 방송 제작 현장의 장벽을 어떻게 허물고 있는지 그 구체적인 여정을 시작해 보고자 한다.

Part 2-1 AI의 귀와 뇌 - 초정밀 음성 인식과 언어의 경계 확장(STT & Diarization)

과거의 음성인식(Speech-to-Text, 이하 STT) 기술이 단순히 '소리를 문자로 받아쓰는' 수준에 머물렀다면, 오늘날 AI 네이티브 방송 환경에서의 STT는 콘텐츠의 구조를 파악하고 제작 공정의 자동화를 완성하는 핵심 엔진으로 진화했다. 방송 엔지니어에게 요구되는 과제는 단순한 엔진 도입을 넘어, 복잡한 제작 현장의 변수를 통제하고 제작진이 즉시 활용할 수 있는 '고신뢰도 텍스트 자산'을 설계하는 것이다.

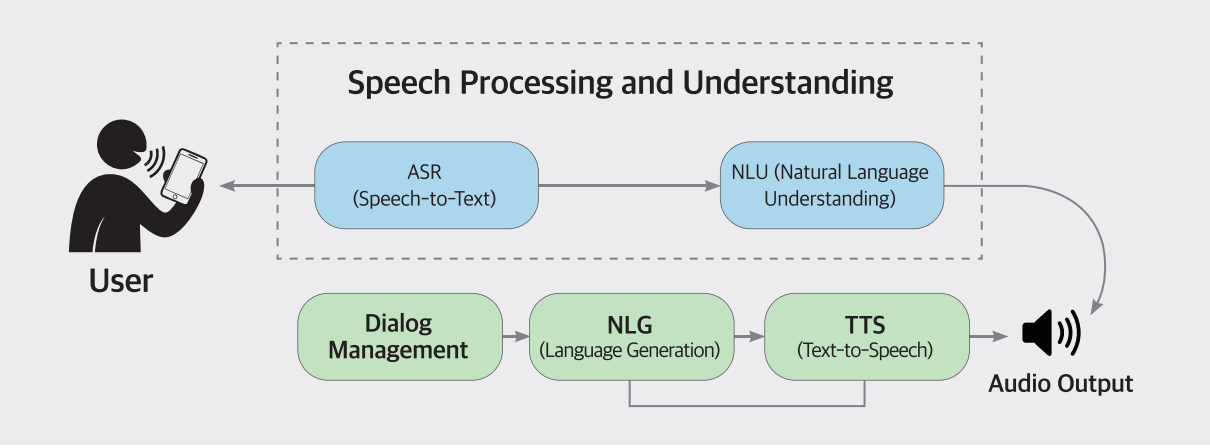

그림 1. 대화형 AI 시스템의 아키텍처 및 흐름도

첫째, 편집 효율을 극대화하기 위한 멀티 화자 분리(Speaker Diarization) 기술을 통한 타임라인의 구조화

정제된 스튜디오에서의 1:1 대담과 달리, 실제 방송 현장은 수많은 출연자가 동시에 발화하고 오디오가 겹치는 복잡한 환경이다. 여기서 AI는 각 화자의 고유한 음성적 특징을 수치화한 '음성 지문(Voice Print)'을 식별한다. 단순히 음성을 텍스트로 바꾸는 것이 아니라, "누가, 언제, 어떤 말을 했는가"를 타임라인별로 분리해내는 것이 핵심이다.

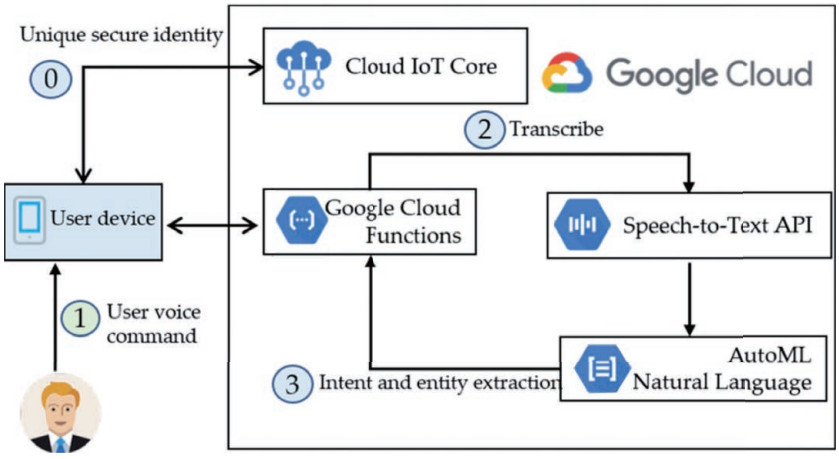

그림 2. Google의 Cloud IoT Core 및 Speech-to-Text API를 활용한 음성 제어 워크플로우 / 출처 : Google cloud speech-to-text platform

기술적 원리는 크게 세 단계로 나뉜다. 오디오 신호에서 유의미한 특징을 추출하는 '특징 추출(Feature Extraction)', 추출된 데이터를 바탕으로 화자가 바뀌는 지점을 찾는 '분할(Segmentation)', 그리고 동일한 음성 지문을 가진 구간을 하나의 화자로 묶는 '클러스터링(Clustering)' 과정이다. 엔지니어는 이 파이프라인을 최적화하여 집단 토크쇼나 돌발 상황이 많은 인터뷰 현장에서도 편집자가 대본을 일일이 대조하지 않아도 되는 '자동화자 태깅' 시스템을 구축할 수 있다. 이는 곧 편집 타임라인 상에서 특정 인물의 발언 구간만을 검색하고 추출할 수 있는 강력한 메타데이터가 된다.

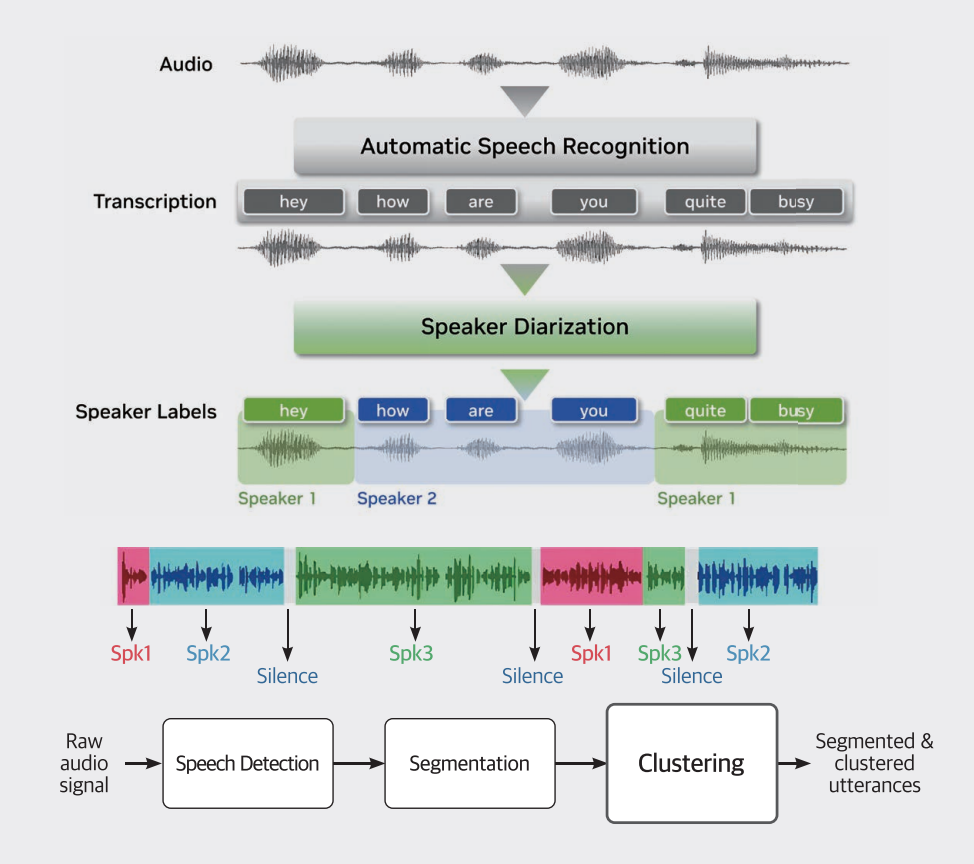

그림 3. Feature Extraction-Segment-Clustering을 통한 화자 식별 프로세스 / 출처 : OpenVINO, www.researchgate.net

둘째, 도메인 특화 언어 모델(Domain-Specific Language Model) 구축을 통한 정확도의 한계 극복

범용적인 AI 모델은 방송 특유의 전문 용어, 고유 인명, 신조어, 혹은 현장 은어 앞에서 무너지는 경우가 많다. 90%의 정확도는 일반 사용자에게 훌륭해 보일지 모르나, 한 글자의 오타도 방송 사고로 이어질 수 있는 제작 현장에서는 99% 이상의 정교함이 요구된다.

이를 위해 엔지니어는 엔진의 '커스텀 사전(Lexicon)'을 관리하고 '미세 조정(Fine-tuning)' 프로세스를 설계해야 한다. 특히 최신 뉴스 텍스트, 보도국 리포트, 프로그램 대본 등 실제 방송 데이터를 집중적으로 학습시켜 모델이 뉴스 도메인의 맥락을 정확히 짚어내도록 만든다.

예를 들어, 일반 AI가 혼동하기 쉬운 정치·경제 분야의 복잡한 인명과 직함, 혹은 뉴스 보도에서 빈번하게 발생하는 전문 용어들을 사전에 학습시켜 인식 오류를 최소화하는 방식이다. 또한 의학 다큐멘터리라면 약학 용어를, 스포츠 중계라면 선수 명단과 기술명을 학습시킨다. 엔지니어는 이 과정에서 데이터의 정제(Cleaning)와 편향성 제거를 관리하며, AI가 '뉴스 보도 및 방송'이라는 특수한 문맥(Context)을 이해하도록 만드는 교육자 역할을 수행한다.

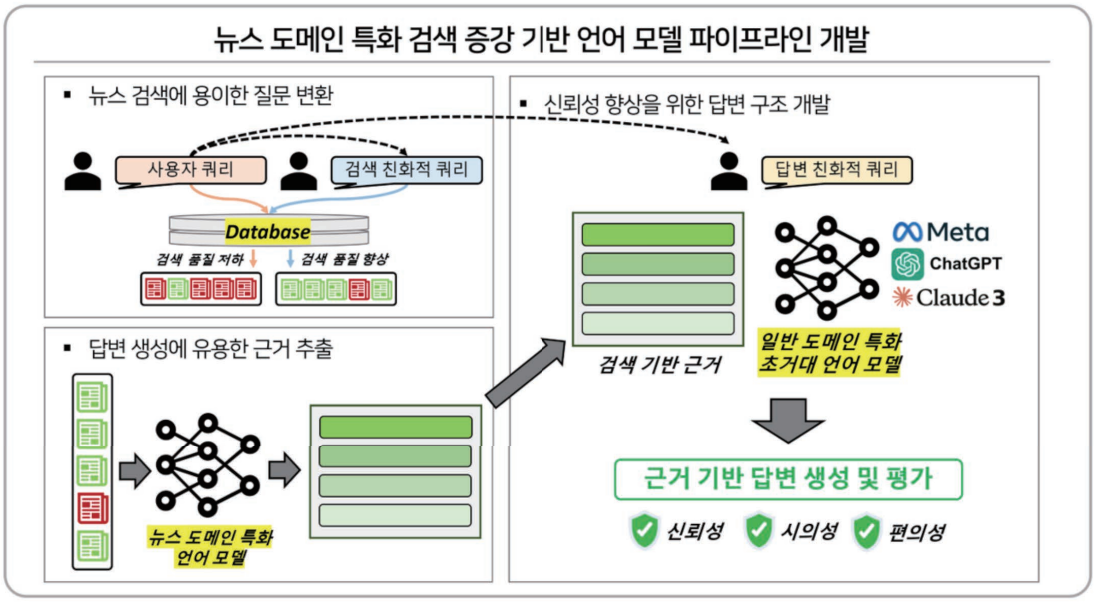

그림 4. 뉴스 도메인 특화 언어모델 사례 / 출처 : dsba.snu.ac.kr

셋째, 다국어 워크플로우 자동화와 제작 툴과의 파이프라인 통합

텍스트가 생성된 이후의 단계는 더욱 공학적이다. 원어로 생성된 STT 결과물은 실시간 AI 번역 엔진과 연동되어 다국어 자막 초안을 즉시 생성한다. 여기서 엔지니어의 핵심 역량은 이 데이터가 '공중에 떠 있지 않게' 만드는 시스템 통합(SI) 능력에 있다.

AI가 생성한 자막 데이터를 프리미어(Adobe Premiere Pro)나 아비드(Avid Media Composer)와 같은 전문 편집 소프트웨어의 자막 트랙(SRT, XML 형식 등)으로 즉시 익스포트(Export) 할 수 있는 자동 파이프라인을 구축해야 한다. 이를 통해 편집자는 자막을 일일이 입력하는 반복 노동에서 해방되어, AI가 만든 초안을 검수하고 예술적인 완성도를 높이는 데에만 집중할 수 있게 된다. 이것이 필자가 강조하는 AI가 주는 '시간의 선물'의 실체다.

☑ 넷플릭스의 글로벌 자막 현지화

AI-Human Hybrid 모델 글로벌 OTT 강자인 넷플릭스는 이 분야에서 가장 앞선 워크플로우를 보여준다. 수십 개국에 동시 배급되는 콘텐츠의 특성상 물리적인 시간 내에 모든 자막을 수동으로 제작하는 것은 불가능하다. 넷플릭스는 AI로 수십 개 언어의 자막 초안을 단시간에 생성한 뒤, 전 세계 전문 검수자들이 플랫폼에 접속하여 최종 수정하는 'AI-Human Hybrid' 시스템을 구축했다.

이 공정에서 기술진은 단순 번역 품질뿐만 아니라, 자막의 글자 수 제한, 화면상의 가독성(Line Break), 음성과의 싱크(Sync) 정확도를 AI가 사전에 검토하도록 아키텍처를 설계했다. 그 결과 넷플릭스는 콘텐츠 배급 속도를 획기적으로 높이면서도 현지화의 품질을 유지하는 데 성공했다. 이는 방송 엔지니어가 단순히 엔진을 돌리는 사람이 아니라, 기술과 인간의 협업 구조를 설계하는 '워크플로우 설계자'임을 증명하는 사례다.

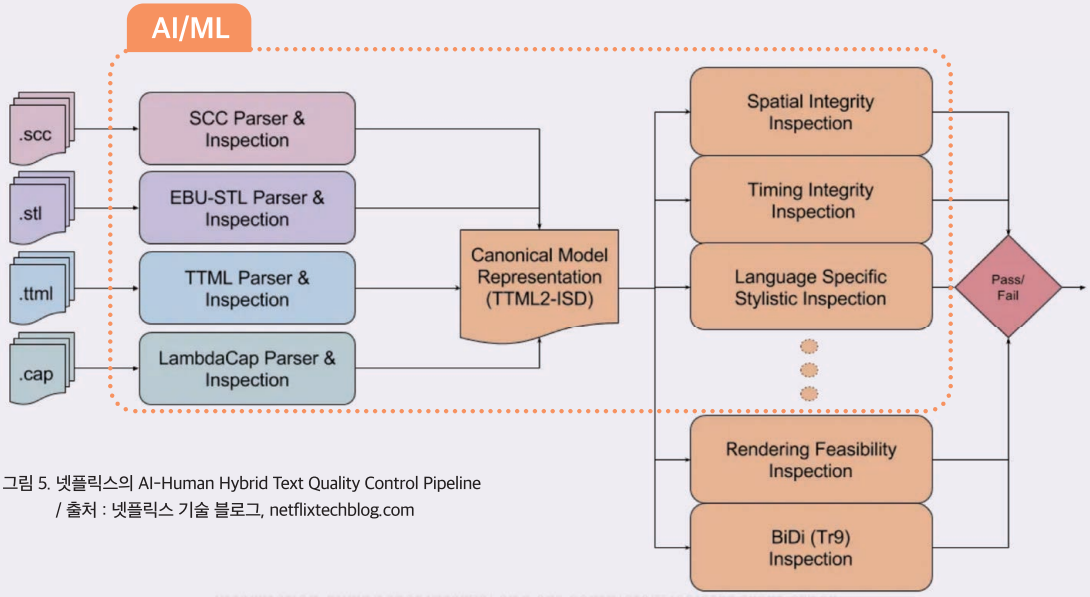

그림 5. 넷플릭스의 AI-Human Hybrid Text Quality Control Pipeline
/ 출처 : 넷플릭스 기술 블로그, netflixtechblog.com

[그림 5]는 넷플릭스가 수천 개의 자막 파일을 효율적으로 검수하기 위해 구축한 '자막 품질 자동화 파이프라인'이다.

- 표준화 과정 : 각기 다른 형식의 자막(.scc, .stl 등)을 하나의 표준 모델(TTML2-ISD)로 변환하여 동일한 기준으로 검사할 수 있게 만든다.
- 다각도 자동 검수 : AI와 알고리즘이 자막의 위치(Spatial), 시간(Timing), 언어적 스타일 등을 동시에 분석하여 기술적 오류를 걸러낸다.
- 효율적 의사결정 : 이 과정을 통해 최종 '통과(Pass)' 또는 '실패(Fail)'를 판정함으로써, 인간 검수자가 핵심적인 수정 작업에만 집중할 수 있도록 돕는다.

이 자동화 시스템 덕분에 넷플릭스는 전 세계 수십 개 언어의 자막을 정확하고 빠르게 배급할 수 있다.

Part 2-2 AI의 귀와 뇌 - 소리의 맥락 해독: 오디오 데이터의 지능형 분석(Audio Intelligence)

음성인식(STT)이 화자의 언어를 텍스트로 치환하는 과정이라면, 오디오 분석(Audio Intelligence)은 대사 이면의 비언어적 요소(배경 소음, 음악적 질감, 발화의 톤)를 해독하여 영상의 맥락을 정의하는 과정이다. AI 네이티브 엔지니어는 소리의 파형에서 의미론적 가치를 추출하여, 아카이브 시스템이 영상의 '분위기'와 '상황'까지 검색할 수 있는 지능형 데이터베이스를 설계해야 한다.

첫째, 엠비언스 및 사운드 이벤트 감지(SED)를 통한 장면의 자동 정의

방송 원본 데이터에는 박수 소리, 환호성, 사이렌, 폭발음 등 수많은 환경음(Ambiance)이 포함되어 있다. 과거에는 이러한 소리들이 그저 오디오 트랙의 일부였으나, 사운드 이벤트 감지(Sound Event Detection) 기술은 이를 통해 장면의 성격을 규정한다. 예를 들어, 스포츠 경기 아카이브에서 '관중의 환호성이 80dB 이상 지속되는 구간'을 AI가 자동으로 태깅하면, 편집자는 별도의 검토 없이도 결정적인 득점 장면이나 극적인 순간을 즉시 찾아낼 수 있다. 엔지니어는 다양한 사운드 라이브러리를 학습시켜 시스템이 긴박한 사이렌 소리나 군중의 비명 등을 식별하게 함으로써, 재난 상황이나 특정 사건 사고 장면을 아카이브에서 초 단위로 소환할 수 있는 검색 인터페이스를 구축할 수 있다.

그림 6. 스포츠 중계의 멀티모달 데이터 구조
/ 출처 : Magnifi /multimodal-ai-sports-content-creation

둘째, 감성 및 톤 분석(Acoustic Analysis)을 통한 콘텐츠의 입체적 자산화

동일한 단어라도 화자의 목소리 톤, 피치(Pitch), 발화 속도에 따라 그 맥락은 완전히 달라진다. 최신 오디오 AI는 음향 특성을 분석하여 화자의 감정 상태(기쁨, 분노, 슬픔, 긴박함 등)를 메타데이터로 추출한다. 엔지니어는 이 데이터를 CMS에 연동하여, 제작진이 '주인공이 분노하며 외치는 장면' 혹은 '차분하고 신뢰감 있는 톤의 내레이션 구간'을 데이터 필드 기반으로 검색하게 만들 수 있다. 이는 단순한 텍스트 검색을 넘어 감성적 맥락에 기반한 영상 편집의 호흡을 제안하는 지능형 편집 어시스턴트의 토대가 된다.

셋째, 오디오 소스 분리(Source Separation)를 통한 제작 유연성의 확보

촬영 현장에서 녹음된 오디오는 대사와 배경음, 음악이 뒤섞인 혼합 신호인 경우가 많다. 엔지니어는 AI 기반의 소스 분리 기술을 활용하여 믹싱된 오디오에서 깨끗한 음성만을 추출하거나, 반대로 음성을 제거하고 배경의 효과음(M&E)만을 남기는 공정을 자동화할 수 있다. 특히 노이즈가 심한 야외 촬영본에서 AI가 주변 소음을 지능적으로 억제하고 화자의 목소리만 복원하는 기술은 후반 작업의 물리적 시간을 획기적으로 단축한다. 엔지니어는 이러한 솔루션을 아키텍처에 통합함으로써 제작진에게 '편집 가능한 깨끗한 소스'를 실시간으로 공급하는 파이프라인을 완성해야 한다.

☑ WSC Sports의 사례

스포츠 중계의 지능형 하이라이트 자동 생성 가장 역동적인 적용 사례는 스포츠 중계 현장에서 찾아볼 수 있다. 글로벌 스포츠 중계권사들은 관중의 함성 크기와 해설자의 음성 톤 변화를 실시간으로 모니터링하는 AI 알고리즘을 운용한다. AI는 경기 중 발생하는 사운드 에너지를 분석하여 골(Goal)이나 결정적인 파울 등 하이라이트가 될 가능성이 높은 지점을 자동으로 감지한다.

여기서 엔지니어의 역할은 단순 분석에 그치지 않고, 감지된 하이라이트 포인트가 즉각적으로 송출용 서버(EVS 등)의 로그로 전송되어 짧은 클립으로 자동 생성되도록 워크플로우를 결합하는 것이다. 경기 종료와 동시에 주요 장면 클립이 SNS나 포털 사이트에 게시될 수 있는 것은, 소리를 '듣는' AI와 이를 '시스템화'한 엔지니어의 설계가 만난 결과다. 결국 오디오 분석 기술은 방송 현장에서 '찰나의 가치'를 실시간 데이터로 변환하여 수익화하는 가장 강력한 도구가 되고 있다.

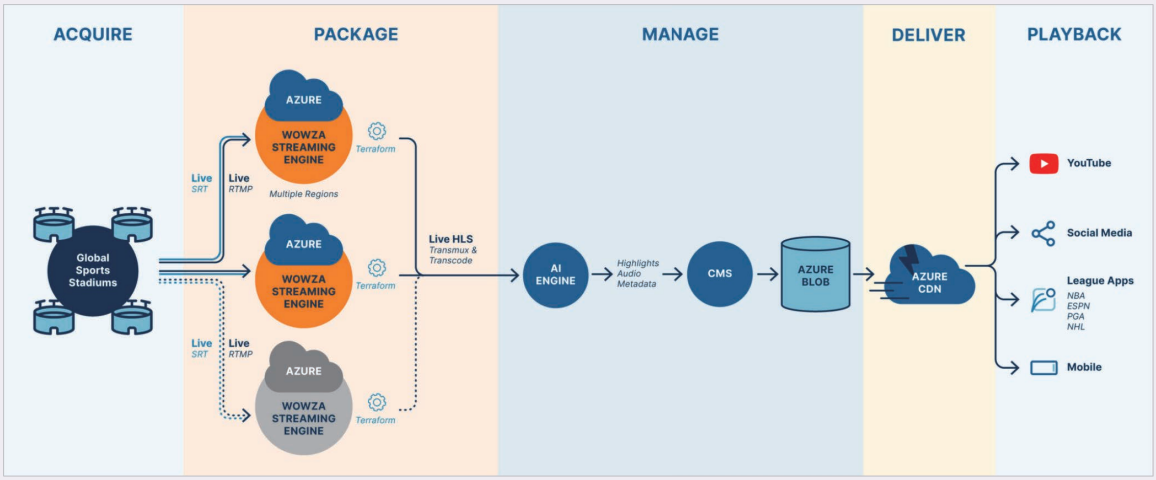

그림 7. AI 기반 지능형 하이라이트 제작 및 오디오 파이프라인 아키텍처 / 출처 : WSC Sports-How it Works

Part 3-1 AI의 입 - 보이스 클로닝과 가상 성우: 고유의 음성 자산 창조(TTS & Voice Cloning)

인공지능이 소리를 '이해'하는 단계를 넘어 '창조'하는 단계에 진입하면서, 방송 제작 현장은 전례 없는 전환점을 맞이하고 있다. 과거의 음성 합성 기술이 딱딱하고 기계적인 어투로 정보를 전달하는 수준에 그쳤다면, 이제는 특정 인물의 고유한 정체성과 감정의 결까지 담아내는 '페르소나 보이스'의 시대가 열린 것이다. 방송 엔지니어에게 이는 단순한 오디오 생성을 넘어, 목소리라는 비정형 자산을 공학적으로 설계하고 관리해야 하는 새로운 임무를 의미한다.

첫째, 뉴럴 TTS(Neural TTS)와 감정 합성 기술을 통한 인간미의 구현

최신 음성 합성 엔진의 핵심은 딥러닝 기반의 뉴럴 네트워크 아키텍처다. 단순히 음소를 이어 붙이던 방식에서 벗어나, AI는 인간 화자의 음조, 리듬, 심지어는 문장 사이의 미세한 호흡까지 학습한다. 특히 '감정 합성' 기술은 발화의 목적에 따라 슬픔, 기쁨, 긴박함 등 상황에 최적화된 목소리를 생성해낸다. 긴급 재난 방송에서는 단호하고 긴박한 톤을, 심야 라디오 프로그램에서는 차분하고 따뜻한 감성을 AI가 스스로 조절하며 출력한다. 엔지니어는 이러한 감정 파라미터를 세밀하게 조정(Tuning)하여, 기계음 특유의 불쾌한 골짜기(Uncanny Valley)를 극복하고 시청자가 거부감 없이 몰입할 수 있는 청각적 환경을 조성해야 한다.

둘째, 보이스 클로닝(Voice Cloning) 기술을 활용한 제작 효율의 극대화

최근의 보이스 클로닝은 단 몇 분, 심지어는 몇 초 분량의 샘플 데이터(Few-shot Learning)만으로도 특정 출연자의 목소리를 정교하게 복제해낸다. 이는 제작 현장에 '시간의 선물'을 안겨주는 핵심 기술이다. 예를 들어, 바쁜 일정의 연예인이 직접 스튜디오에 방문하지 않더라도 가이드 녹음이나 간단한 내레이션을 AI로 대체할 수 있다. 엔지니어는 출연자의 동의하에 확보된 음성 데이터를 안전하게 자산화하고, 이를 퓨샷(Few-shot) 학습 모델에 적용하여 급박한 제작 일정 속에

서도 고품질의 오디오 소스를 실시간으로 생성해내는 파이프라인을 구축한다. 이는 출연진의 물리적 제약을 극복하고 후반 작업의 유연성을 극대화하는 공학적 해법이 된다.

셋째, 가상 성우 아키텍처 설계와 다국어 워크플로우의 통합

엔지니어는 성우의 컨디션이나 장소의 제약 없이 24시간 가동 가능한 '가상 성우 시스템'을 제작 워크플로우에 심어야 한다. 텍스트 수정만으로 별도의 재녹음 없이 즉시 오디오를 갱신할 수 있는 이 시스템은 뉴스 속보나 정보성 프로그램에서 막대한 위력을 발휘한다. 더 나아가, 하나의 원천 목소리를 기반으로 다국어 TTS 엔진과 결합하면 한국어 화자의 목소리 톤을 그대로 유지한 채 영어, 일본어, 스페인어 내레이션을 생성하는 것도 가능하다. 엔지니어는 이러한 다국어 보이스 합성 엔진을 MAM 및 송출 시스템과 유기적으로 통합하여, 글로벌 배급을 위한 현지화 작업을 자동화하고 제작 비용을 획기적으로 절감하는 아키텍처를 완성해야 한다.

☑ 고인(故人)의 목소리 복원 프로젝트

기술과 감동의 접점 지능형 음성 기술이 대중에게 가장 큰 울림을 준 사례는 바로 세상을 떠난 인물의 목소리를 복원한 프로젝트들이다. 방송기술진은 고인이 남긴 과거의 저품질 아카이브 데이터를 수집하여 배경 소음을 제거하고 깨끗한 음성 데이터셋을 구축한다. 이후 이를 딥러닝 모델에 학습시켜 고인 특유의 말투와 발성 습관을 완벽하게 재현해낸다.
실제로 다큐멘터리나 특집 예능에서 구현된 고인의 목소리는 시청자들에게 정서적 위로와 깊은 감동을 선사했다. 하지만 엔지니어는 여기서 기술적 성취에만 매몰되어서는 안 된다. 고인의 목소리를 활용함에 있어 유가족의 동의와 윤리적 가이드라인 준수는 필수적이며, AI가 생성한 목소리임을 명확히 고지하는 기술적 장치(Watermarking 등) 역시 엔지니어가 책임져야 할 영역이다. 이는 기술이 인간의 존엄성과 공존하며 가치를 발현하는 대표적인 사례로, 엔지니어가 단순한 기술자를 넘어 '가치 전달의 매개자'임을 보여준다.

그림 8. AI 기반 故김광석 복원 프로젝트 방송 사례 / 출처 : SBS <세기의 대결 AI vs 인간>

Part 3-2 AI의 입 - 오디오 복원과 업스케일링: 시간의 먼지를 닦아내는 기술(Audio Restoration)

지능형 음성 기술이 가져온 또 하나의 혁명적 변화는 '시간의 제약'을 극복하는 복원력에 있다. 방송 아카이브에 잠들어 있는 수십 년 전의 소리들은 테이프의 열화, 녹음 환경의 한계, 그리고 당시 기술의 미비함으로 인해 현대의 고화질 영상 콘텐츠와 결합하기에는 품질이 턱없이 부족한 경우가 많다. AI 네이티브 엔지니어는 인공지능을 활용해 오디오 데이터에 쌓인 '시간의 먼지'를 닦아내고, 과거의 유산을 현재의 방송 규격으로 업스케일링하는 심폐소생술을 집도해야 한다.

첫째, 지능형 노이즈 제거(AI De-noising)를 통한 음성 명료도의 혁신

과거의 노이즈 제거 기술이 특정 주파수 대역을 일괄적으로 깎아내어 목소리까지 왜곡시켰다면, AI 기반의 디노이징은 딥러닝 모델이 '인간의 목소리'와 '소음'을 실시간으로 분리(Speech Separation)해낸다. 촬영 현장의 강한 바람 소리, 자동차 엔진음, 혹은 열악한 실내의 울림(Reverb) 속에서도 AI는 화자의 음성 파형만을 정교하게 보존하고 나머지 불필요한 신호는 완벽에 가깝게 제거한다. 엔지니어는 이러한 솔루션을 편집 아키텍처에 통합하여, 현장 녹음 상태가 좋지 않은 소스라도 별도의 재녹음(ADR) 없이 스튜디오급 품질로 개선함으로써 제작진에게 물리적인 제작 시간의 단축이라는 '시간의 선물'을 안겨준다.

둘째, 오디오 대역폭 확장(Bandwidth Extension) 기술을 활용한 청각적 업스케일링

오래된 아카이브 영상이나 저해상도 녹음본은 주파수 대역폭이 좁아 소리가 답답하고 뭉개지는 특성을 보인다. 엔지니어는 생성형 AI 모델을 통해 소실된 고주파수 대역을 추론하여 복원하는 대역폭 확장 기술을 적용한다. 이는 마치 저해상도 사진을 고해상도로 바꾸는 것과 같은 원리로, 전화 통화 수준의 좁은 음역대를 하이파이(Hi-Fi)급의 풍성한 음향으로 탈바꿈시킨다. 엔지니어는 이 과정을 통해 시청자가 과거의 영상을 보면서도 청각적으로는 이질감 없이 현대적인 몰입감을 느낄 수 있도록 오디오의 품질을 상향 평준화하는 역할을 수행한다.

셋째, 스피치 인핸스먼트(Speech Enhancement)와 디지털 리마스터링의 자동화

방송 현장에서는 화자가 중얼거리듯 발음하거나 마이크와의 거리가 멀어 대사 전달력이 떨어지는 경우가 빈번하게 발생한다. AI는 발음의 명료도를 높이고 화자의 음성 에너지를 최적화하는 스피치 인핸스먼트 공정을 수행하여, 자막 의존도를 낮추고 콘텐츠의 전달력을 극대화한다. 엔지니어는 이러한 개별 기술들을 하나의 '디지털 리마스터링 워크플로우'로 규격화하여, 방대한 양의 아카이브 데이터를 대량으로 정제하고 현대적인 표준 라우드니스(Loudness) 규격에 맞게 자동 보정하는 시스템을 설계해야 한다.

☑ 아카이브 음원 현대화 프로젝트

과거와 현재를 잇는 공학적 가치 실제 방송 현장에서는 수십 년 전의 필름이나 오픈 릴 테이프를 디지털화하는 과정에서 AI 오디오 복원 기술이 적극적으로 도입되고 있다. 과거의 특집 다큐멘터리나 역사적 인터뷰 영상을 복원할 때, 지능형 복원 엔진은 지직거리는 히스(Hiss) 노이즈와 테이프 씹힘 현상으로 인한 왜곡을 실시간으로 감지하고 메운다.

특히, 4K나 8K로 업스케일링된 고화질 영상에 맞춰 오디오 역시 5.1 채널이나 돌비 애트모스(Dolby Atmos)급의 입체 음향으로 확장하는 공정은 엔지니어의 창의적 엔지니어링이 빛을 발하는 지점이다. 기술을 통해 과거의 소리를 현대의 시청자에게 생생하게 전달하는 이 작업은, 단순히 음질을 개선하는 것을 넘어 잊혀 가던 기록에 생명력을 불어넣는 사회적 가치를 지닌다. 결국 오디오 복원 기술은 방송기술자가 기술을 통해 시대를 연결하고, 시간이라는 장벽을 허물 수 있음을 증명하는 가장 강력한 수단이다.

그림 9. SKT, AI로 복원한 '815 리마스터링' 사례
/ 출처 : SK Telecom Newsroom

2024년 SK텔레콤이 자체 인공지능(AI) 기술을 이용해 1945년 광복 전후의 영상 및 음원을 더욱 선명하게 복원하는 디지털 프로젝트를 진행했다. 이름하여 '815 리마스터링'이다. SKT는 79번째 광복절을 맞아, 광복 직후의 풍경을 생생하게

복원하고 많은 이들이 그날의 감격적 순간을 간접 경험해보도록 돕자는 취지에서 프로젝트를 기획했었다. 프로젝트에 쓰인 콘텐츠 원본은 1945년 광복 직후 서울 거리 영상, 그리고 1942년 녹음된 애국가 음원이다. 개선된 두 콘텐츠를 합해 새롭게 제작한 '815 리마스터링' 영상은 SKT 공식 유튜브 채널의 시리즈 [AI help you?]에서 시청할 수 있다. 이때, 복원한 영상 원본은 광복 직후인 1945년 8월 16일 서울 거리의 만세 행렬 등을 담은 28초짜리 자료다. 8월 15일 광복 사실을 몰랐던 많은 사람이 하루 뒤인 16일에야 거리로 쏟아져 나왔는데, 그 장면이 담겨있다.

독립기념관 관계자는 "선조들이 독립의 각오를 다지며 불렀던 애국가 음원과 광복 당시 영상이 오늘날의 AI 기술을 만나 개선된 콘텐츠로 복원될 수 있어 뜻깊게 생각한다"며 "더욱이 내년 광복 80주년을 앞두고, 자료에 담긴 선열들의 독립정신과 광복 당시 환희가 담긴 영상을 통해 광복의 감격을 선명하게 기억할 수 있게 된 건 매우 의미있다"고 말했다고 한다.

Part 4 결론: 소리로 완성하는 미디어 지능화의 퍼즐

지금까지 우리는 AI의 '입과 귀'를 통해 소리라는 비정형 데이터가 어떻게 방송 제작 현장의 핵심 자산으로 거듭나는지 그 공학적 여정을 살펴보았다. 지난 호에서 다룬 컴퓨터 비전(CV) 기술이 영상 속 객체와 인물을 식별하며 시각 정보의 무결성을 확보했다면, 이번 호에서 탐구한 지능형 음성 기술은 그 시각 정보에 '맥락'과 '의도'라는 깊이를 더했다. 결국 미디어 지능화의 완성은 눈과 귀가 각각의 데이터를 분석하는 단계를 넘어, 이 둘이 유기적으로 결합하여 콘텐츠를 입체적으로 이해하는 '멀티모달(Multimodal)'의 단계에서 비로소 이루어진다.

음성인식(STT)을 통해 추출된 텍스트 메타데이터는 단순히 자막 제작 시간을 단축하는 도구에 그치지 않는다. 이는 시각 분석 데이터와 결합하여 '특정 인물이 특정 단어를 언급하며 분노하는 장면'과 같은 고차원적인 검색을 가능케 하며, 방송 엔지니어가 '신호의 흐름'을 넘어 '의미의 흐름'을 설계하는 데이터 전략가로 진화해야 함을 시사한다. 또한, TTS와 보이스 클로닝을 통한 음성 생성 기술은 제작진에게 물리적 시공간을 초월한 창작의 자유를 부여하며, 인공지능이 인간의 창의성을 해방하는 '시간의 선물'로서의 역할을 수행하고 있음을 확인했다.

이제 방송 엔지니어의 임무는 더욱 명확해졌다. 우리는 단순히 최신 AI 엔진을 도입하는 운영자를 넘어, 분석된 원시 데이터(Raw Data)를 SMPTE나 EBU와 같은 국제 표준 메타데이터로 가공하고 이를 MAM이나 CMS 등 기존 방송 인프라와 완벽하게 통합하는 시스템 아키텍트가 되어야 한다. AI가 추출한 '로고 위치'와 '대사 내용'이 구체적인 좌표와 타임코드를 가진 '상품화된 메타데이터'로 변환될 때, 비로소 데이터는 제작자가 즉시 활용할 수 있는 가치 있는 자산이 된다.
AI의 눈이 영상을 창조하고, AI의 입이 소리를 지배하며 제작 자동화의 기반을 닦았다면, 이제 우리는 이 방대한 지능형 자산을 어떻게 시청자에게 최적으로 전달하고 수익화할 것인가라는 다음 단계의 과제를 마주하게 된다. 지능형 음성 기술이 안겨준 제작 공정의 혁신은 결코 종착역이 아니다. 이는 콘텐츠가 가진 본질적인 가치를 극대화하고, 시청자 개개인에게 가장 정교한 방식으로 다가가기 위한 필수적인 과정이다. 기술이 창조를 빛나게 하는 무대를 설계하는 지휘자로서, 엔지니어는 AI라는 강력한 도구를 우리 현장에 맞게 끊임없이 길들여야 한다. 소리로 완성된 미디어 지능화의 퍼즐 조각은 이제 사용자 데이터와 만나 '초개인화'라는 더 큰 그림을 그려나갈 것이다.

다음 6회차에서는 AI가 단순한 추천을 넘어 어떻게 시청자의 취향을 정교하게 큐레이션하고, 광고 공학과의 결합을 통해 미디어 산업의 새로운 수익 모델을 제시하는지 'AI 큐레이터의 등장 - 초개인화 추천과 광고의 공학'이라는 주제로 심층적으로 다루고자 한다.