전자통신동향분석, 40권 6호 2025, 24-34

시청각 음성인식 기술 동향 분석

Trends in Audiovisual Speech Recognition Technology

박기영 (K.Y. Park, pkyoung@etri.re.kr) 체화복합지능연구실 책임연구원
강점자 (J.J. Kang, jjkang@etri.re.kr) 체화복합지능연구실 책임기술원
오창한 (C.H. Oh, ochh508@etri.re.kr) 체화복합지능연구실 UST학생연구원
최우용 (W.Y. Choi, wychoi4@etri.re.kr) 체화복합지능연구실 책임연구원
한란 (R. Han, ran.han@etri.re.kr) 체화복합지능연구실 선임연구원
송화전 (H.J. Song, songhj@etri.re.kr) 체화복합지능연구실 책임연구원/실장

ABSTRACT

Audiovisual Speech Recognition (AVSR) jointly analyzes speech audio and lip movements to overcome the limitations of audio-based recognition in noisy or multispeaker environments. This paper reviews the concepts and necessity of AVSR, research trends, benchmark datasets, key technological advances, market outlook, and policy issues. Visual cues help compensate for weaknesses in audio recognition, enabling applications such as real-time captions for the hearing-impaired, voice commands in vehicles, improved captions at video conferences, and security or surveillance systems. End-to-end deep learning has driven rapid progress since 2016, with large-scale self-supervised learning resulting in additional improvements. Recent work has explored deeper visual context integration, robustness in multispeaker and noisy scenarios, and the combination of AVSR with large language models. Challenges remain in terms of privacy, large-scale data acquisition, standardization, and environmental robustness.

KEYWORDS Automatic Speech Recognition, Audio-visual Speech Recognition, Multi-modal AI

I. 서론

시청각 음성인식(AVSR: Audiovisual Speech Recognition) 기술은 사람의 청각과 시각을 동시에 활용하는 음성인식 기술을 뜻한다. 사람은 대화할 때 상대방의 목소리뿐만 아니라 입술의 움직임, 표정 등의 시각적 단서도 함께 이용한다. 예를 들어, 시끄러운 환경에서도 상대의 입 모양을 보면 말을 더 잘 이해

* DOI: https://doi.org/10.22648/ETRI.2025.J.400603
* This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT)(RS-2022-II220989(2022-0-00989), Development of Artificial Intelligence Technology for Multi-speaker Dialog Modeling).

할 수 있는데, 이러한 원리를 기계 음성인식에 적용한 것이 AVSR 기술이다. 다시 말해, 마이크로 소리 신호를 입력받는 기존 자동음성인식(ASR: Automatic Speech Recognition)에 카메라로 포착한 화자의 입술 움직임 영상을 결합하여 인식하는 것이다.

이러한 멀티모달 접근은 자연스러운 인간 의사소통 방식을 모방하기 때문에, 음성인식의 정확도와 신뢰성을 높일 수 있다는 장점이 있다. 음향 신호는 배경소음이나 반향 등에 영향받아 왜곡될 수 있고, 영상 신호는 조명이나 화자의 얼굴 움직임에 따라 인식이 어려워질 수 있다. 하지만 오디오와 비디오 정보의 결합을 통해 서로의 약점을 보완하면, 훨씬 강건한 음성 인식이 가능해진다. 실제 인간 청각 연구에서도 "맥거크 효과(McGurk Effect)"라고 하는 현상이 알려져 있는데, 이는 동일한 음성이라도 입 모양 영상에 따라 다르게 들릴 정도로 시각 정보가 음성 이해에 중요함을 보여준다[1]. 이런 인간의 특성을 고려할 때, 시청각 음성인식 기술은 필연적 발전 방향이라고 볼 수 있다.

최근 인공지능 기술의 발달로 이러한 멀티모달 처리 능력이 크게 향상됨에 따라 AVSR은 음성인식 분야의 새로운 패러다임으로 떠오르고 있다. 다음 장에서는 왜 지금 AVSR 기술이 주목받고 있으며, 어떠한 기술적 진전이 있었는지 살펴본다.

II. 기술의 필요성 및 배경

왜 AVSR이 필요한가? 첫째, 기존의 오디오 기반 음성인식 시스템은 조용한 실내 등 통제된 환경에서는 높은 성능을 보이지만, 현실 세계의 소음 환경에서는 인식률 저하 문제가 있다. 거리의 소음, 여러 사람이 동시에 말하는 상황, 음향 반향이 심한 공간 등에서는 마이크로 포착한 음성 신호만으로 정확한 인식이 어렵다. 반면 사람은 입 모양과 같은 시각 정보를 활용함으로써 이런 상황에서도 의미를 파악할 수 있다. 따라서 기계도 사람처럼 멀티모달 통합을 하면, 예측 성능을 개선할 수 있다는 기대가 있다.

둘째, 청각 장애인 지원 측면에서 AVSR 기술의 필요성이 크다. 청각 장애인은 소리를 듣기 어려우므로 입술 읽기에 의존해 의사소통하는 경우가 많은데, 사람마다 발음과 입 모양이 달라 정확히 읽기 어렵다. 만약 AI가 화자의 음성과 입 모양을 동시에 인식하여 실시간 자막이나 텍스트 변환을 제공해준다면, 청각 장애인의 소통을 혁신적으로 도울 수 있다. 실제로 2016년 영국 옥스퍼드대학교 연구팀이 개발한 립넷(LipNet)은 입술 영상만으로도 문장을 읽어내는 소프트웨어로, 93.4%의 판독 정확도를 달성하여 화제가 되었다[2]. 이는 전문 독순술가의 정확도 52%를 훨씬 뛰어넘는 수준으로, 사람보다 정확하게 입 모양만 보고 말을 이해하는 AI의 등장을 알린 사건이었다. 이처럼 AVSR 기술은 장애인 보조 기술로서도 큰 잠재력을 지닌다.

셋째, 일상에서 AI 비서나 음성으로 동작하는 기기가 늘어나면서, 더 자연스럽고 편리한 인터페이스에 대한 요구가 커지고 있다. 마이크와 카메라가 함께 장착된 스마트폰, 스마트 스피커, 자동차 등이 등장함에 따라, 멀티모달 인터페이스가 현실적으로 가능해졌다. 예를 들어, 밤에 조용한 방에서 스마트폰 음성비서에게 명령을 내릴 때 목소리를 내기 어렵다면, 카메라 앞에서 입 모양만 조용히 움직여도 명령을 인식하는 기능이 있다면 유용할 것이다. 최근에는 Physical AI에 대한 관심 확대로 휴머노이드 로봇 시장이 성장할 것으로 전망된다. 이때 로봇과의 소통은 인간 간 대화처럼 시청각을 포함한 멀티모달 기반으로 확장될 가능성이 크다. 특히 어레이 마이크와 영상 정보를 결합하면 극심한 소음 환경에서도 안정적인 인식이 가능하다. 이러한 수요에

대응해 AVSR 기술은 자연스러운 차세대 인터페이스로 주목받고 있다.

이러한 필요성들이 맞물려, 최근 AVSR 기술에 관한 관심과 연구가 폭발적으로 증가하고 있다. 다음 장에서는 기술적으로 어떤 발전이 있었기에 주목받게 되었는지, 주요 연구 동향과 혁신 사례를 살펴본다.

III. 기술 발전의 배경 및 트렌드

1. 딥러닝 이전의 초기 연구

시청각 음성인식의 아이디어 자체는 새로운 것이 아니다. 이미 1990년대 후반부터 음성인식 성능 향상을 위해 비디오 영상정보를 결합하려는 연구가 시작되었다[3,4]. 당시에는 주로 화자의 입술 주변 영역을 추적하여 손수 설계한 특징(예: 입술 모양의 너비, 높이, 윤곽 등)을 추출하고, 이를 음성인식기에 함께 입력하는 방식이었다. 몇몇 연구에서 이러한 오디오-비디오 결합 시스템이 음성만 사용하는 경우보다 오류율을 낮추는 성과를 보이면서, 멀티모달 인식의 시너지 효과에 대한 가능성이 입증되었다[5]. 사람의 입 모양을 읽는 독순 기술을 컴퓨터에 구현한다는 의미에서 이 분야를 자동 독순(Automatic Lip-Reading)이라고도 부른다.

다만, 초기의 AVSR 시도들은 기술적 한계로 인해 널리 실용화되지는 못했다. 그 당시에는 컴퓨터 비전 기술이 미흡하여 영상에서 입술을 정확히 추적하기도 어려웠고, 추출한 시각 특징을 음성 인식 모델에 융합하는 방법도 제한적이었다. 또한, 음향-영상 데이터를 대량으로 확보하기도 힘들어서 주로 실험실 환경의 소규모 데이터로 성능을 검증하는 수준에 머물렀다. 그 결과, 초기 AVSR 시스템은 특정 문장이나 단어에 한정된 실험이 많았고, 실생활의 복잡한 언어를 인식하기에는 역부족이었다. 이러한 상황이 획기적으로 바뀌게 된 계기가 2010년대 들어 등장한다.

2. 딥러닝과 데이터 시대의 도래

2010년대 중반, 딥러닝 혁명과 함께 음성인식 기술은 질적으로 도약했다. 마이크로소프트, 구글 등에서 딥러닝 기반의 대규모 음성인식 모델을 도입한 이후 사람 수준에 근접하는 정확도가 달성되었고[6], AI 스피커와 음성비서가 상용화되기 시작했다[7,8]. 이와 발맞추어 영상 인식 분야도 딥러닝 기반의 합성곱신경망(CNN), 순환신경망(RNN) 등의 발전으로 얼굴 인식, 객체인식 성능이 비약적으로 향상되었다[9]. 이러한 기술 환경의 변화는 오디오와 비디오를 함께 학습할 수 있는 토대를 마련했다. 복잡한 특징 공학(Feature Engineering)에 의존하지 않고도, 대량의 훈련 데이터만 있다면 신경망이 최적의 멀티모달 특징을 자동으로 학습할 수 있게 된 것이다.

특히 2016년은 AVSR 분야의 중요한 전환점으로 평가된다. 옥스퍼드대학교의 LipNet은 딥러닝을 활용해 연속된 문장 단위의 자동 독순을 처음으로 성공적으로 시연하며 주목받았다[2]. 같은 해, 옥스퍼드대학교 연구진은 대규모 자연발화 문장 수준의 LRS2 데이터셋을 공개해 이후 딥러닝 기반 AVSR 연구의 사실상 표준 벤치마크를 마련했다[10]. LipNet은 제한된 문장과 화자 환경에서도 93% 이상의 정확도를 기록하며 기존 인간 독순 능력을 크게 앞질렀다. 두 연구는 각각 자연 환경 대화의 대규모 데이터 구축과 제한된 환경에서의 고정밀 독순이라는 서로 다른 성과를 거두었지만, 공통적으로 딥러닝 기반 AVSR의 실용 가능성을 전 세계에 각인시킨 이정표로 평가된다.

이후로 AVSR 기술은 학계와 산업계 양쪽에서 급물살을 탔다. 주요 딥러닝 학회에서는 멀티모달 음

성 처리 세션이 등장했고, 음성인식 챌린지에도 오디오+비디오 부문이 신설되는 등 연구자들의 관심이 높아졌다[11]. 기술적으로는 두 개의 신경망(하나는 음향 특성을, 하나는 영상 특성을 학습)으로 구성된 병렬 구조 모델, 또는 음성 및 영상 입력을 한꺼번에 처리하는 멀티모달 트랜스포머 모델 등이 시도되었다[12,13,33]. 또한, Attention 메커니즘을 통해 시간 축 상에서 음성과 영상 정보를 효과적으로 융합하는 기법들이 개발되었다[10]. 그 결과 잡음 환경에서의 단어 오류율이 크게 떨어지고, 일부 한정된 테스트에서는 AVSR이 오디오 단일 모델보다 50% 이상 오류율을 감소시키는 등 유의미한 향상이 보고되었다[12].

더욱 최근의 흐름은 대규모 자기지도학습(Self-Supervised Learning)과 사전학습된 거대 음성 · 비전 모델의 결합이다. 대표적으로 메타(META)가 제안한 AV-HuBERT는 정답 레이블이 없는 대규모 데이터를 마스크 예측(가려진 구간 예측) 방식으로 사전 학습하여, 소량의 라벨만으로도 최고 수준의 성능을 보였다[14]. 데이터 라벨링 병목을 줄이려는 시도로는 임페리얼 칼리지 팀의 Auto-AVSR도 있다. 대규모 비라벨 영상에서 자동 생성한 자막을 활용해 연속 문장 AVSR을 학습하는 접근으로, 실제 라벨 데이터의 사용을 최소화하면서도 강건한 성능을 달성했다[15]. 또한, RAVEn은 모멘텀 티처-스튜던트 구조와 마스킹 예측을 결합해 오디오 · 비디오 공동 표현을 한 단계에서 학습하는 자기지도 프레임워크로, 저라벨 및 고라벨 양쪽 설정에서 LRS3 기준 SOTA에 도달하며 완전 원시 데이터에서의 양질 표현 학습 가능성을 입증했다[16].

사전학습 음성모델과 시각 표현을 결합한 Whisper-Flamingo 계열도 주목받는다. OpenAI의 대규모 ASR Whisper 디코더에 Flamingo식 시각 크로스-어텐션을 삽입해 입술 특징을 주입하는 방식으로, 잡음 환경에서 오디오 단일 모델 대비 유의한 개선을 보였다[17]. 더 나아가 mWhisper-Flamingo는 AV-HuBERT 기반 비디오 인코더와 Whisper를 결합해 다국어(9개 언어) AVSR에서 잡음 강건성을 확대했으며, 다국어 세팅에서도 오디오 전용 Whisper보다 일관되게 우수한 성능을 보고했다[18]. 이처럼 자기지도 대규모 사전학습 모델의 멀티모달 결합은 라벨 절감, 잡음 강건성, 다국어 확장성 측면에서 AVSR의 실용화를 가속하고 있다.

3. 최신 동향

최근 AVSR 분야의 가장 두드러진 흐름 중 하나는 대규모 언어모델(LLM)을 결합한 멀티모달 학습이다. 기존의 자기지도학습이 라벨 없이 음성 · 영상 표현을 학습하는 데 집중했다면, 이제는 LLM의 강력한 언어 이해와 추론 능력을 활용해 오디오 · 비주얼 정보를 통합적으로 처리하려는 시도가 활발하다. 대표적인 연구로는 LLaMA-AVSR과 MMS-LLaMA가 있다[19,20].

이들 모델은 Meta의 LLaMA 계열 LLM에 오디오 · 비디오 멀티모달 인코더를 결합하여, 음성과 영상 특징을 LLM의 문맥 처리 능력 속에 통합한다. 특히 LLaMA-AVSR은 사전학습된 음성 · 영상 인코더로부터 추출한 특징을 LLaMA 입력 공간으로 매핑하는 방식을 채택한다. 인코더와 LLM 본체는 동결한 채, 연결 모듈만 경량 학습함으로써 효율성과 성능을 동시에 확보하였다. 반면 MMS-LLaMA는 조기 오디오 · 비디오 융합과 동적 쿼리 할당 방식을 도입하여 입력 길이나 발화 속도에 따라 멀티모달 특징을 효율적으로 처리한다. 이를 통해 불필요한 계산을 줄이면서도 높은 인식 성능을 유지하는 것이 특징이다.

표 1 공개 시청각(Audiovisual) 데이터셋 목록 및 요약

데이터셋	환경/출처	언어	발화수	화자수	발화 내용
AVLetters[21]	실험실	영어	780	10	영어 알파벳 단일 발화
CUAVE[22]	방음부스	영어	7.9k	36	개별 및 연속 숫자 발화
GRID[23]	방음부스	영어	34k	34	간단한 명령문
AV-TIMIT[24]	실험실	영어	4.4k	223	음소 균형을 맞춘 짧은 문장
OuluVS[25]	실험실	영어	817	20	짧은 구문
OuluVS2[26]	스튜디오	영어	1.5k	53	구문 및 문장
AVICAR[27]	자동차 내부	영어	59k	86	구문 및 문장
OLKAVS[28]	스튜디오	한국어	2.5M	1,107	읽기 문장
LRW[10]	방송뉴스	영어	539k	>1,000	비정형 자발적 문장 속의 단일 단어
LRS2[12]	방송뉴스	영어	144k	-	비정형 자발적 문장
LRS3-TED[29]	온라인강연	영어	152k	9,545	비정형 자발적 문장
VoxCeleb1[30]	유튜브영상	영어	154k	1,251	비정형 자발적 문장(전사 없음)
VoxCeleb2[31]	유튜브영상	다국어	1.1M	6,112	비정형 자발적 문장(전사 없음)
KMSAV[32]	유튜브영상	한국어	>43.8k	-	비정형 자발적 문장

출처 Reproduced from K.Y. Park et al., “KMSAV: Korean multi-speaker spontaneous audiovisual dataset,” ETRI J., vol. 46, no. 1, 2024, pp. 71-81.

IV. AVSR 데이터셋

AVSR 연구를 본격적으로 진행하기 위해서는 신뢰할 수 있는 대규모 데이터셋 확보가 무엇보다 중요하다. 최근 성능 향상은 모델 구조의 혁신뿐만 아니라 다양한 환경 · 화자 · 언어를 아우르는 공개 데이터셋의 등장 덕분에 가능했다. 본 장에서는 이러한 배경을 토대로 언어권별 또는 환경별로 널리 활용되는 대표적 데이터셋을 살펴본다. 이는 연구자들이 모델 성능을 공정하게 비교하고 자기지도학습이나 다국어 확장 등 최신 기법을 검증하는 데 필요한 기반을 제공하며, 향후 한국어 AVSR 연구 및 상용화를 위한 데이터 인프라 전략에도 중요한 시사점을 준다.

표 1은 현재 AVSR 연구에서 활용되는 주요 데이터셋을 요약한 것이다. 다음 절에서는 이들 중 대표적인 데이터셋 몇 가지를 선정해 특징과 구성, 그리고 활용사례를 보다 자세히 소개한다.

1. 영어권 AVSR 데이터셋

1.1 LRW: Lip Reading in the Wild

옥스퍼드대학교와 BBC가 2016년에 발표한 대규모 단어 단위 영상/음성 데이터셋으로, BBC 뉴스 영상에서 추출한 자연발화 단어 클립 500,000개 이상으로, 1,000개 단어 어휘를 포함하며 총 57시간 분량에 달한다[10]. 배경이 다양하고 현실 뉴스 장면이라 자연스러운 구어체 발음과 다양한 화자가 포함된다. 단어 수준 시각인식(립리딩) 및 AVSR 성능을 크게 끌어올렸으며, LRW-1000 등 다른 언어 대규모 립리딩 데이터 구축의 계기가 되었다.

1.2 LRS2-BBC & LRS3-TED

옥스퍼드대학교는 2017~2018년 사이에 문장 단위 대규모 AVSR 데이터셋 시리즈인 LRS2와 LRS3를 공개했다. LRS2는 BBC 방송에서 추출한 자연 문장을 포함하고 있으며, 수만 개의 발화를 담은 데이

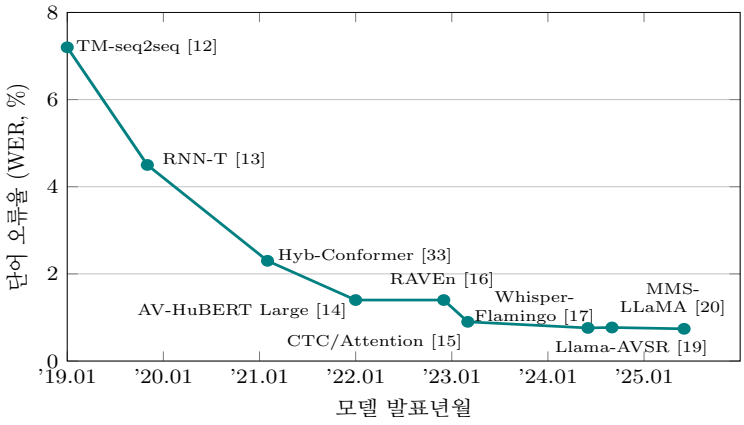

그림 1 LRS3 데이터셋을 이용한 주요 음성인식 모델들의 발표 시기별 단어 오류율(WER) 감소 경향

터셋이다[12]. LRS3는 TED/TEDx 영상에서 추출한 400시간 이상의 비디오 및 대화 클립을 포함하고 있으며, 5,594개의 강연 데이터를 기반으로 제작되었다[29]. 두 데이터셋 모두 제한 없는 어휘와 자연발화를 포함하고 있어, 대어휘 연속 AVSR 연구의 표준 벤치마크로 널리 사용된다. 특히 LRS3는 다양한 억양과 발음을 포함하므로 난이도가 더 높다. 옥스퍼드대학교 연구진은 이 LRS 데이터를 활용해 트랜스포머 기반 AVSR 모델을 훈련시켜 유의미한 성능 향상을 보였고, 이후 많은 문장 단위 AVSR 연구가 LRS2/3를 기준으로 평가된다. 그림 1은 LRS3-TED 데이터셋을 이용하여 수행된 연구들의 성능을 시간 순서대로 나열한 것이다.

1.3 VoxCeleb2

옥스퍼드대학교가 2018년에 발표한 대규모 연예인/유튜버 영상 음성 데이터셋으로, 6,000여 명의 다국적 화자가 포함된 2,442시간 분량 영상으로 구성되나, 발화 내용 정답 전사문이 없어 화자 인식(Speaker Recognition) 용도로 주로 쓰인다[31]. AVSR 연구에서는 VoxCeleb의 방대한 다국어 음성-영상 데이터를 비지도학습 형태의 사전훈련에 활용하기도 한다.

2. 한국어 AVSR 데이터셋

2.1 OLKAV: Open Large-scale Korean Audio-Visual Speech

대규모의 공개 한국어 시청각 음성 데이터셋으로 일반인 1,107명이 참가해 스튜디오에서 일상 문장을 읽은 오디오 1,150시간 분량에, 각각 9가지 카메라 각도 영상(총 5,750시간)을 동시 녹화하여 다각도 정보를 제공한다. 또한, 실내 · 실외, 무소음 · 생활소음 · 도로소음 등 여러 조건에서 녹음되어 청정 및 잡음 환경을 망라한다. 한국정보화진흥원 AI-Hub에 공개되어 누구나 활용 가능하며, 제공된 베이스라인 모델로 멀티뷰 학습 시 단일뷰 대비 오류 감소 등 시청각+다각도 학습 효과를 검증하였다[28].

2.2 KMSAV: Korean Multi-speaker Spontaneous Audio-Visual Speech

한국전자통신연구원은 더욱 현실적인 조건의 한국어 AVSR 연구를 위해 YouTube에서 다화자가 대화하는 영상을 수집하고 정제해 공개했다. 이 데이터셋은 약 150시간 분량의 전사 및 주석된 시청

각 자료와 함께, 크리에이티브 커먼즈 라이선스로 확보된 2,000시간 이상의 비전사 영상을 포함한다[32]. 해당 데이터는 연구 목적에 제약 없이 무료로 제공되며, ASR 및 AVSR용 오픈소스 프레임워크도 배포된다. 최신 ASR 및 AVSR 기법을 적용해 미세조정 기반 평가를 수행했으며, 조정 후 ASR과 AVSR은 각각 문자 오류율(CER) 11.1%, 18.9%를 기록했다. 이 같은 성능 차이는 AVSR 기술의 여전한 개선 여지를 강조하는 지표로 작용한다.

V. AVSR 기술의 주요 활용 분야

시청각 음성인식 기술은 사람의 말을 인식하는 모든 분야에 잠재적인 응용이 존재한다. 그 중 특히 주목받는 활용 분야들을 몇 가지 살펴보면 다음과 같다.

1. 청각 장애인 보조 및 접근성

AVSR의 사회적으로 가장 의미 있는 활용은 청각 장애인을 위한 의사소통 보조 기기이다. 상대방이 말할 때 스마트폰이나 안경형 디바이스의 카메라로 입 모양을 인식하고, 이를 문자로 변환하여 실시간 자막을 띄워준다면 청각 장애인의 소통 장벽을 크게 낮출 수 있다. 예를 들어, 회의나 수업 시간에 자동 자막을 제공하는데 음성만으로 부족한 경우, 화자의 입술 영상까지 활용하면 자막 정확도를 높이고 대화의 뉘앙스까지 더 잘 전달할 수 있다. 또한, AVSR 기술은 발음 교정 도구로도 활용될 수 있는데, 사용자가 소리를 내며 말할 때 AI가 그의 입 모양과 음성을 동시에 분석하여 올바른 발음 여부를 피드백해주는 것이다. 이는 청각 장애인뿐만 아니라 일반인들의 언어 학습, 언어치료에도 쓰일 수 있는 기술이다.

2. 키오스크 및 공공 단말기

공공장소의 무인 키오스크나 안내 단말기 환경에서도 AVSR은 매우 유용한 기술로 주목받고 있다. 음식점의 무인 주문기, 지하철 역사 내 안내 단말기, 병원 접수 · 안내 시스템 등은 주변 소음이 큰 경우가 많아 전통적인 음성인식만으로는 정확도가 떨어질 수 있다. 이때 카메라로 사용자의 입 모양을 함께 인식하면 음성만으로는 놓치기 쉬운 발화를 보완할 수 있다. 예를 들어, 식당 키오스크에서 고객이 메뉴를 말로 주문할 때 주방 소음이나 배경 음악으로 인해 음성이 잘 들리지 않더라도, AVSR은 입 모양 정보를 활용해 보다 정확하게 명령을 인식할 수 있다. 또한, 고령층이나 외국인 관광객 등 발음이나 억양이 다양한 사용자에게도 접근성과 사용 편의성을 높이는 핵심 기술로 활용될 수 있다.

3. 휴머노이드 로봇

최근 Physical AI에 대한 관심이 커지면서 휴머노이드 로봇 시장이 빠르게 성장할 것으로 전망된다. 로봇과의 소통은 단순한 음성 명령을 넘어, 사람과 사람 사이의 대화처럼 시청각을 포함한 멀티모달 기반으로 확장될 필요가 있다. 예를 들어, 가정 내 서비스 로봇이나 산업 현장의 안내 로봇이 소음을 동반한 환경에서도 안정적으로 음성을 인식하려면, 음성 신호뿐만 아니라 사용자의 입 모양을 동시에 해석하는 AVSR 기술이 필수적이다. 또한, 어레이 마이크와 카메라를 결합하면 극심한 소음 환경에서도 인식 정확도를 크게 높일 수 있다. 실제로 일부 연구에서는 이러한 접근법이 잡음이 많은 조건에서 로봇의 대화 이해 능력을 향상시키는 데 효과적임을 확인한 바 있으며, 이는 향후 휴머노이드 로봇의 상용화 과정에서 중요한 기반 기술로 작용할

것이다.

4. 스마트폰 및 가상비서

스마트폰, 스마트 스피커, AI 비서 등 일반 소비자 기기에도 AVSR의 활용이 논의된다. 현재도 음성인식 기반의 음성비서(예: Bixby, Siri)가 널리 쓰이고 있지만, 소음 환경이나 조용히 말해야 하는 환경에서는 제약이 있다. AVSR을 적용하면 사용자가 굳이 큰 소리로 말하지 않아도 카메라 앞에서 조용히 입술만 움직여서 명령을 내릴 수 있게 될 전망이다. 예컨대 도서관에서 스마트폰에 검색 명령을 내리거나, 집에서 모두 잠든 밤에 AI 스피커에 알람 설정을 할 때 목소리를 내지 않고도 실행시킬 수 있다. 이는 프라이버시 면에서도 유용하다. 또한, 향후 AR 글라스와 연동하여 입 모양 기반으로 텍스트 입력하거나 메뉴를 제어하는 등 새로운 HCI 방식도 등장할 수 있다.

5. 원격회의 · 방송 및 고객상담

코로나19 이후 화상회의, 비대면 방송 등이 일상화되면서 자동 자막 생성 기술의 중요성이 커졌다. 줌(Zoom)이나 구글 미트(Meet) 같은 회의 플랫폼은 이미 실시간 음성 자막 기능을 제공하지만, 음성만으로는 종종 오인식이 발생한다. 여기에 화자의 영상(특히 입술)까지 분석하는 AVSR이 더해지면 자막 정확도를 높이고, 화자가 누구인지 구별하는 화자 분리도 더 정확해질 수 있다. 예를 들어, 동시에 두 사람이 말할 때 입 모양을 보면 누가 무슨 말을 했는지 분리인식이 가능하다. 방송 분야에서도 스튜디오 토론 프로그램 등에 AVSR을 적용해 자막을 생성하거나, 나아가 방송콘텐츠를 자동으로 다국어 더빙하거나 수어 아바타 생성에 활용하는 연구도 진행 중이다. 한편, 은행 창구나 콜센터의 영상 상담 서비스에서 고객의 말이 잘 들리지 않을 때를 대비해 카메라로 입 모양을 읽어 상담사가 놓친 내용을 보완하는 식의 활용도 생각해 볼 수 있다. 이는 고객 서비스 품질 향상과 커뮤니케이션 장애 극복 차원에서 의미 있는 응용이 될 것이다.

6. 보안 및 법 집행

AVSR 기술은 보안 분야에도 특별한 가치를 지닌다. 사용자 인증 시 음성+얼굴 정보를 함께 사용하면 안전성을 높일 수 있다. 예컨대 스마트폰 잠금해제나 전화금융 서비스 본인확인 시 사용자의 얼굴이 “안녕하세요, 홍길동입니다.”라고 말하는 영상+음성 이중 인증을 요구하면, 음성만 녹음으로 위조하거나 얼굴만 딥페이크로 합성해서는 뚫기 어려운 강력한 보안이 된다. 이밖에 수사기관에서는 용의자의 CCTV 영상에 녹음된 음성이 없는 경우에도 입술 판독으로 대화를 복원하려는 시도가 이루어지고 있다. 현재는 화질과 각도 문제로 제한적이나, 앞으로 기술이 발전하면 저해상도 영상에서도 일정 수준 대화 내용을 추론해 내는 것이 가능해질 것으로 보인다. 다만 이런 활용은 개인 생활 침해 논란을 불러일으킬 수 있어, 기술 윤리와 정책적 논의가 병행되어야 할 민감한 분야이다.

이 외에도 군사통신(소음 많은 전장에서 입 모양으로 지휘 전달), 비밀통화(타인이 못 듣게 입술로만 명령), 영화 및 콘텐츠 제작(무성 영화 복원이나 영상 속 대사 자동추출) 등 아이디어에 따라 다양한 분야에서 AVSR 기술을 접목할 수 있다. 핵심은 사람이 말하고 듣는 모든 상황에 시각 정보를 추가하면 더 풍부한 소통이 가능하다는 점이다. 그렇기에 AVSR의 응용 범위는 앞으로 계속 확대될 것으로 전망된다.

VI. 시장 전망 및 주요 이슈

1. 시장 성장 전망

전 세계 음성인식 기술 시장은 AI 기술 발전과 함께 지속하여 성장하고 있으며, AVSR은 이 시장의 차세대 핵심 기술로 기대를 모은다. 시장조사기관 Fortune Business Insights에 따르면 2024년 전 세계 음성 및 음성인식(Speech and Voice Recognition) 시장 규모는 약 154억 달러로 추산되며, 2032년에는 816억 달러에 이를 전망이다[34]. 이는 2025년부터 2032년까지 연평균 23.1%에 달하는 매우 높은 성장률이다. 이러한 성장의 주요 동인은 고객센터의 음성 IVR 확대, 스마트폰 및 가전의 음성 인터페이스 채택, 차량용 음성비서 등이다.

AVSR 기술 자체만의 시장 규모를 산정한 자료는 아직 드물지만, 업계에서는 멀티모달 통합이 차세대 음성인식의 표준이 될 것으로 보고 있다. 다시 말해, 향후 음성인식 솔루션에는 영상 인식 기능이 포함되는 것이 경쟁력으로 작용할 가능성이 크다. 실제로 음성 AI 선도기업들은 방대한 딥러닝 모델을 구축하면서 오디오뿐만 아니라 비디오까지 활용한 고성능 모델을 개발하려는 움직임을 보인다. 이렇게 되면 음성인식 시장의 파이가 커질 뿐만 아니라 기존 음성인식이 진출하지 못했던 영역에도 새로운 진입이 가능해진다. 예를 들어, 극심한 소음 현장(공장, 공사장 등)이나 수화기 너머로 음성만 들리던 콜센터에도 AVSR을 통한 혁신의 여지가 생기는 것이다.

2. 고려해야 할 주요 이슈

시장 성장 가능성과 별개로, AVSR 기술이 대중화되기 위해서는 몇 가지 해결해야 할 이슈들이 존재한다.

2.1 프라이버시와 윤리

음성인식과 달리 영상 정보를 수집하는 AVSR은 사용자 프라이버시 침해에 대한 우려를 높일 수 있다. 항상 카메라로 얼굴을 촬영해야 한다는 점에서 사용자들은 사생활 침해나 감시당하는 느낌을 받을 수 있다. 예를 들어, 가정용 AI 스피커에 카메라가 달려 있다면 거부감을 줄 수 있으며, 직장에서도 직원 컴퓨터에 상시 카메라로 입 모양을 감시하는 것은 수용되기 어렵다. 따라서 영상 데이터의 처리와 보안에 대한 신뢰 확보가 필수적이다. 수집된 영상이 외부로 유출되거나 다른 목적으로 악용되지 않도록 기기 내 온-디바이스 처리를 강화하고, 얼굴 영상 자체를 저장하지 않고 필요한 특징만 추출 후 폐기하는 등의 개인정보보호 설계가 요구된다. 이와 함께 AVSR 기술의 악용, 예컨대 CCTV로 상대 몰래 대화를 엿듣는 등의 행위를 금지할 법적 · 윤리적 기준 마련도 중요하다.

2.2 데이터 확보와 학습 효율

딥러닝 기반 AVSR 시스템을 개발하려면 대량의 병렬 음성-영상 데이터(사람이 말할 때의 소리와 그 영상)가 필요하다. 하지만 이런 데이터는 언어별로 구하기 어렵고, 라벨링(정답 문자 변환)에도 시간이 많이 든다. 특히 한국어처럼 음소 단위 입 모양 변화가 미묘한 언어의 경우, 다양한 화자의 발화 영상을 확보하고 주석 달기하는 작업이 만만치 않다. 앞서 언급했듯 AI Hub 등에서 국가 차원의 데이터셋 공개가 이뤄지고 있으나, 현재 공개된 5천여 시간은 영어권에 비해 부족한 실정이다. 또한, 영상 데이터는 용량이 커서 학습 비용이 많이 들고, 처리 속도도 느릴 수 있다. 이를 극복하려면 효율적인 모델 구조와 증강기술(Augmentation), 전이학습(Transfer Learning) 등의 기법 연구가 계속되어야 한다. 최근 자기지도 학습으로 적은 레이블 데이터로도 높은 성능을 내

는 방향이 고무적이지만, 여전히 사람 수준의 인식을 위해서는 수만 시간 이상의 데이터가 필요하다.

2.3 기술적 한계

현재의 AVSR 모델들도 완벽하지 않다. 화자가 카메라를 정면으로 바라보지 않거나, 얼굴 일부가 가려지면 성능이 급격히 떨어진다. 예를 들어, 마스크를 착용하면 입술 영상을 쓸 수 없고, 옆모습만 보이면 정확한 입 모양을 얻기 어렵다. 조명이 어둡거나 카메라 해상도가 낮은 환경, 혹은 화자의 발화 속도가 너무 빠르거나 억양이 강한 경우에도 인식률 저하가 발생한다. 사람은 이런 상황에서도 추론을 통하여 어느 정도 알아듣지만, 기계는 약한 모습을 보인다. 따라서 어려운 조건에서도 강인한 AVSR을 만들기 위한 연구가 필요하다. 가능하다면 적외선 카메라를 사용해 어두운 곳에서도 입술을 추적한다든지, 3D 센서를 통해 옆모습에서도 입술 형상을 추정하는 등의 보조기술도 고려할 수 있다. 또한, 여러 명이 동시에 말할 때 화자별로 영상을 구분해 각각의 음성을 식별해내는 다화자 음성인식 문제와 결합하면 난도가 더욱 높아지는데, 이 역시 풀어야 할 과제다.

요약하면, AVSR 기술이 시장에서 본격적으로 꽃피우기 위해서는 기술 신뢰성 확보와 사용자 수용성 제고라는 두 측면의 과제를 해결해야 한다. 이는 비단 기술 개발자뿐만 아니라 정책 입안자, 산업계 모두의 협력이 필요한 부분이다.

VII. 결론

시청각 음성인식(AVSR)은 인간의 청각과 시각을 모방하여 음성인식의 한계를 극복하려는 차세대 인공지능 기술로, 최근 딥러닝과 자기지도학습, 그리고 대규모 언어모델(LLM)과의 결합을 통해 빠르게 진화하고 있다. 특히 잡음 환경, 다화자 대화, 프라이버시 친화적 온디바이스 처리 등 현실적 요구를 해결하기 위한 연구가 활발히 진행 중이다.

앞으로 AVSR은 청각 장애인 보조, 휴머노이드 로봇 인터페이스, 스마트 단말기, 보안 및 법 집행 등 다양한 분야에서 실질적 가치를 창출할 것으로 기대된다. 동시에 개인정보 보호, 윤리적 가이드라인, 사회적 수용성 확보와 같은 과제도 병행되어야 한다.

결론적으로, AVSR은 단순한 음성인식을 넘어 멀티모달 인간-기계 소통의 핵심 기반 기술로 자리 잡을 것이며, 연구 · 산업계의 협력과 사회적 합의가 더해질 때 우리 삶 속에서 안전하고 신뢰할 수 있는 기술로 확산될 것이다.

참고문헌

[1] H. McGurk and J. MacDonald, “Hearing lips and seeing voices,” Nature, vol. 264, 1976, pp. 746-748.

[2] Y.M. Assael et al., “LipNet: End-to-End Sentence-level Lipreading,” arXiv preprint, 2016. doi: 10.48550/arXiv.1611.01599

[3] G. Potamianos et al., “Recent Advances in the Automatic Recognition of Audiovisual Speech,” Proc. IEEE, vol. 91, no. 9, 2003, pp. 1306-1326.

[4] C. Neti et al., “Audio-Visual Speech Recognition,” Final Workshop 2000 Report, Center for Language and Speech Processing, Johns Hopkins University, Oct. 2000.

[5] K. Saenko et al., “Visual Speech Recognition with Loosely Synchronized Feature Streams,” in Proc. IEEE Int. Conf. Comput. Vis. (Beijing, China), Dec. 2005, pp. 1424-1432.

[6] L. Deng et al., “Recent Advances in Deep Learning for Speech Recognition at Microsoft,” in Proc. IEEE Workshop Spoken Lang. Technol., (Vancouver, BC, Canada), May. 2013, pp. 33-38.

[7] Siri Team, “Hey Siri: An On-device DNN-powered Voice Trigger for Apple’s Personal Assistant,” Apple Machine Learning

Research Blog, 2017. 10. 1. https://machinelearning.apple.com/research/hey-siri

[8] A. Rastrow and S. Mevawalla, "On-device speech processing makes Alexa faster, lower-bandwidth," Amazon Science Blog, 2022. 1. 25.

[9] L. Liu et al., "Deep Learning for Generic Object Detection: A Survey," Int. J. Comput. Vis., vol. 128, no. 2, 2020, pp. 261-318.

[10] J.S. Chung et al., "Lip Reading Sentences in the Wild," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., (Honolulu, HI, USA), Jul. 2017, pp. 3444-3453.

[11] H. Chen et al. "The First Multimodal Information Based Speech Processing (Misp) Challenge: Data, Tasks, Baselines And Results," in Proc. IEEE Int. Conf. Acoust. Speech Signal Process., (Singapore, Singapore), Apr. 2022, pp. 9266-9270.

[12] T. Afouras et al., "Deep Audio-Visual Speech Recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 12, 2022, pp. 8717-8727.

[13] T. Makino et al., "Recurrent Neural Network Transducer for Audio-Visual Speech Recognition," in Proc. IEEE Autom. Speech Recognit. Understanding, (Sentosa, Singapore), Dec. 2019, pp. 905-912.

[14] B. Shi et al., "Robust Self-Supervised Audio-Visual Speech Recognition," arXiv preprint, 2022. doi: 10.48550/arXiv:2201.01763

[15] P. Ma et al., "Auto-AVSR: Audio-Visual Speech Recognition with Automatic Labels," in Proc. IEEE Int. Conf. Acoust. Speech Signal Process., (Rhodes Island, Greece), 2023, pp. 1-5.

[16] A. Haliassos et al., "Jointly Learning Visual and Auditory Speech Representations from Raw Data (RAVEn)," in Proc. Int. Conf. Learn. Represent, (Kigali, Rwanda), May. 2023.

[17] A. Rouditchenko et al., "Whisper-Flamingo: Integrating Vsisual Features into Whisper for Audio-Visual Speech Recognition and Translation," in Proc. Interspeech, (Kos, Greece), Sep. 2024, pp. 2420-2424.

[18] A. Rouditchenko et al., "mWhisper-Flamingo for Multilingual Audio-Visual Noise-Robust Speech Recognition," arXiv preprint, 2025. doi: 10.48550/arXiv:2502.01547

[19] U. Cappellazzo et al., "Large Language Models are Strong Audio-Visual Speech Recognition Learners," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process, (Hyderabad, India), 2025, pp. 1-5.

[20] J.H. Yeo et al., "MMS-LLaMA: Efficient LLM-based Audio-Visual Speech Recognition with Minimal Multimodal Speech Tokens," in Proc. Comput. Linguist., (Vienna, Austria), Jul. 2025, pp. 20724-20735.

[21] I. Matthews et al., "Extraction of visual features for lipreading," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 2, 2002, pp. 198-213.

[22] E.K. Patterson et al., "CUAVE: A new audio-visual database for multimodal human-computer interface research," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., (Orlando, FL, USA), May. 2002, pp. II-2017-II-2020.

[23] M. Cooke et al., "An audio-visual corpus for speech perception and automatic speech recognition," J. Acoust. Soc. Am., vol. 120, no. 5, 2006, pp. 2421-2424.

[24] T.J. Hazen et al., "A Segment-Based Audio-Visual Speech Recognizer: Data Collection, Development, and Initial Experiments," in Proc. Int. Conf. Multimodal Interfaces, (State College, PA, USA), Oct. 2004, pp. 235-242.

[25] G. Zhao et al., "Lipreading With Local Spatiotemporal Descriptors," IEEE Trans. Multimedia, vol. 11, no. 7, 2009, pp. 1254-1265.

[26] I. Anina et al., "OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis," in Proc. IEEE Int. Conf. Automatic Face Gesture Recognit., (Ljubljana, Slovenia), Jul. 2015, pp. 1-5.

[27] B.W. Lee et al., "AVICAR: audio-visual speech corpus in a car environment," in Proc. Interspeech, (Jeju Island, Rep. of Korea), Oct. 2004, pp. 2489-2492.

[28] J.K. Park et al., "OLKAVS: An Open Large-Scale Korean Audio-Visual Speech Dataset," in Proc. IEEE Int. Conf. Acoust. Speech Signal Process., (Seoul, Rep. of Korea), Apr. 2024, pp. 6385-6389.

[29] T. Afouras et al., "LRS3-TED: a large-scale dataset for visual speech recognition," arXiv preprint, 2018. doi: 10.48550/arXiv:1809.00496

[30] A. Nagrani et al., "VoxCeleb: Large-scale speaker verification in the wild," Comput. Speech Lang., vol. 60, 2020, p. 101027.

[31] J.S. Chung et al., "VoxCeleb2: Deep Speaker Recognition," in Proc. Interspeech, (Hyderabad, India), Sep. 2018, pp. 1086-1090.

[32] K.Y. Park et al., "KMSAV: Korean multi-speaker spontaneous audiovisual dataset," ETRI J., vol. 46, no. 1, 2024, pp. 71-81.

[33] P. Ma et al., "End-To-End Audio-Visual Speech Recognition with Conformers," in Proc. IEEE Int. Conf. Acoust. Speech Signal Process., (Toronto, ON, Canada), Jun. 2021, pp. 7613-7617.

[34] Fortune Business Insights, "Speech and Voice Recognition Market to Reach $81.59 Billion by 2032," 2025. 4. 1. https://www.fortunebusinessinsights.com/press-release/speech-and-voice-recognition-market-9266