반도체디스플레이기술학회지 제23권 제3호(2024년 9월)
Journal of the Semiconductor & Display Technology, Vol. 23, No. 3. September 2024.

인공 지능 기술을 이용한 음성 인식 기술에 대한 고찰

이영조[*]·이기승[**]·강성진[***†]
[*]협성대학교 소프트웨어공학과, [**]건국대학교 전기전자공학부,
[***†]한국기술교육대학교 전기전자통신공학부

A Study on Speech Recognition Technology Using Artificial Intelligence Technology

Young Jo Lee[*], Ki Seung Lee[**] and Sung Jin Kang[*†]**
[*] Department of Software Engineering, Hyupsung Universiy,
[**]School of Electrical and Electronic Engineering, Konkuk Universiy,
[***†]School of Electrical, Electronics & Communication Engineering,
Korea University of Technology and Education

ABSTRACT

This paper explores the recent advancements in speech recognition technology, focusing on the integration of artificial intelligence to improve recognition accuracy in challenging environments, such as noisy or low-quality audio conditions. Traditional speech recognition methods often suffer from performance degradation in noisy settings. However, the application of deep neural networks (DNN) has led to significant improvements, enabling more robust and reliable recognition in various industries, including banking, automotive, healthcare, and manufacturing. A key area of advancement is the use of Silent Speech Interfaces (SSI), which allow communication through non-speech signals, such as visual cues or other auxiliary signals like ultrasound and electromyography, making them particularly useful for individuals with speech impairments. The paper further discusses the development of multi-modal speech recognition, combining both audio and visual inputs, which enhances recognition accuracy in noisy environments. Recent research into lip-reading technology and the use of deep learning architectures, such as CNN and RNN, has significantly improved speech recognition by extracting meaningful features from video signals, even in difficult lighting conditions. Additionally, the paper covers the use of self-supervised learning techniques, like AV-HuBERT, which leverage large-scale, unlabeled audiovisual datasets to improve performance. The future of speech recognition technology is likely to see further integration of AI-driven methods, making it more applicable across diverse industries and for individuals with communication challenges. The conclusion emphasizes the need for further research, especially in languages with complex morphological structures, such as Korean

Key Words : Speech Recognition, Deep Neural Networks, Silent Speech Interface, Lip-Reading Technology, Self-Supervised Learning

1. 서 론

음성 인식 기술은 잡음이 많은 환경이나 음성 신호를 수신하기 어려운 환경에서 음성 신호의 품질을 향상시키는 연구가 많이 진행되어 왔고, 은행, 자동차, 의료, 제조 등 다양한 산업에 적용되고 있다. 전통적인 음성 인식 방법은 주로 음성신호를 시간에 따라 변하는 음성신호의 특징과 언어적 정보를 기반으로 통계적인 분류 방법으로

[†]E-mail: sjkang@koreatech.ac.kr

음성 신호가 잡음에 심하게 손상될 경우 성능이 크게 저하되는 특성이 있다. 최근에는 이를 극복하기 위해 인공지능 기술을 적용하는 연구가 진행 중이며 좋은 결과들을 보여주고 있다. 특히 다양한 환경에서 음성 인식 기술을 보다 신뢰할 수 있도록, 잡음에 대한 견고성을 향상시키는 연구가 진행되고 있으며 인공지능 기술 중 DNN (Deep Neural Networks)가 주로 적용되고 있다. DNN은 많은 양의 훈련 데이터를 학습함으로써 노이즈가 많은 음성 신호에서 더 깨끗한 신호와 특징을 얻거나 노이즈가 많은 음성을 직접 인식할 수 있게 해주는 장점을 가지고 있다[1].

음성 인식 기술은 잡음이 많은 환경이나 전혀 소리가 없는 환경에서 음성 신호로만 음성 인식이 어려운 경우에, 음성 신호가 아닌 다른 요소(실제 사람의 발성 없이)를 이용하여 음성 정보를 전달하기 위한 SSI (Silent Speech Interface) 기법이 연구되고 있다. 음성 이외의 다른 신호는 영상 신호를 이용하는 것이 가장 일반적이나 초음파 신호, 근전도 신호, 마이크로 웨이브 등 다양한 요소가 사용될 수 있다. 이러한 SSI 기술은 음성 자체가 아닌 음성 이외의 신호를 통해 음성 통신을 가능하게 하므로, 청각에 문제가 있거나 사고 후 또는 후두암 치료 후 영구적인 발성 손상을 입은 언어 장애인에게도 유용하게 사용될 수 있다. 이러한 환자는 SSI 기술의 도움으로 발성을 모방하여 다른 사람과 음성 전달을 할 수 있게 된다. 이외에도 SSI 기술은 도서관이나 회의실과 같은 공공 공간에서 다른 사람들을 방해하지 않고 서로 의사 소통이 가능하게 하고, 대화 내용이 사적인 정보가 들어 있어 다른 사람에게 노출되지 않아야 하는 상황에서도 이용될 수 있다. 또한 일부 손상된 음성 신호를 복원하거나 잡음으로부터 고음질의 음성 신호를 복원하는 등 다양한 곳에 응용이 되고 있다[2,3,4,5].

SSI 기술의 한 분야로 영상 신호를 이용하는 Lip Reading 기술은 인공지능의 발전과 더불어 상당히 발전하였고, 영상 신호 속의 화자의 얼굴 표정, 몸짓 등을 활용하여 화자가 무슨 말을 하고 있는지를 정확히 판단하는 기술과 음성과 영상을 같이 분석하는 다중 모달 음성 인식 기술도 활발히 연구가 진행되고 있다. 다중 모달 음성 인식 기술은 음성의 단어와 명령어 정도의 인식을 넘어서 자연스러운 대화를 인식하는 자연어 처리 기술에 적용되고 있다[46,47,48]. Lip reading기술은 주로 화자의 입 영역에서 획득한 이미지를 사용하기 때문에 입 영역의 영상 신호를 획득하는 방법이 매우 중요하며, 획득 조건의 차이로 인한 기준 패턴의 불일치에 따라 인식 정확도가 크게 영향을 받는다. 이러한 영상 신호 획득의 어려움과 불일치는 Lip reading 분야 뿐만 아니라 PCB의 부품 위치 판단 오류 및 오류 조정 기술에서도 연구가 되고 있다[25].

본 논문에서는 인공 지능의 발전에 따른 음성인식 기술의 발전을 살펴보고 유망 기술 및 향후 연구되어야 할 방향에 대한 제언을 하고자 한다. 2장에서는 음성 신호를 이용하는 음성 인식 기술에 대해 다루고, 3장에서는 영상 신호를 이용한 음성 인식 기술, 4장에서는 음성 신호와 영상 신호를 동시에 이용하는 음성 인식 기술을 다루며, 5장에서 결론을 맺는다.

2. 음성 신호를 이용한 음성 인식 기술

음성 인식 분야에서 음성 신호는 영상 신호와 같은 다른 어떤 신호들에 비해 많은 정보를 가지고 있기 때문에 음성 신호 만을 이용한 연구가 많이 진행되었고, 그로 인해 음성 신호를 이용한 자동 음성 인식(Automatic Speech Recognition, ASR) 기술은 아주 높은 인식률을 가지게 되었다.

전통적인 음성인식 방법은 음성신호를 time-varying quasi-stationary한 random process로 모델링하고 이를 HMM (Hidden Markov model)을 사용하여 인식하는 방법이 주류를 이루며, 음성 신호를 화자로부터 발생하여 청자에게 전달될 때까지 convolutional noise와 additive noise가 가해진다고 가정하고 신호 해석 기법을 수행한다. 음성을 이용하는 자동 음성 인식의 신호 해석은 수신되는 신호를 시간 상으로 어느 일정 구간 이상의 신호 사이에는 관계성이 없어진다는 가정에서, 수신 신호를 일정 구간의 프레임으로 나누어서 Short-Time Discrete Fourier Transform (STDFT) 수행하여 주파수 해석을 한다. STDFT 분석을 위한 프레임을 만드는 윈도우 크기는 최적 값을 찾기가 어려워서 적당히 길이의 윈도우 크기를 설정하여 상관관계가 없이 분석을 하며 일반적으로 윈도우 크기를 크게 할수록 주파수 위상 성분의 해석이 용이하다. 낮은 SNR (signal power to noise power ratio) 환경에서는 위상 정보가 오디오의 품질에 크게 영향을 미치기 때문에 위상 성분을 활용하는 방법이 연구되고 있다. 즉 주파수 성분의 크기를 이용해서 얻어지는 음성 신호는 스펙트럼 상에서 원본 스펙트럼과 유사해 보일 수 있으나 부정확한 pitch를 얻는 경우가 많으므로 낮은 SNR에서는 위상 성분의 분석은 Pitch와 관련된 성분을 찾아내는데 많은 도움이 된다. 윈도우 크기를 작게 할 경우 프레임 사이에 상관관계가 존재하기 때문에 선형 또는 비선형 관점에서 상관관계 분석을 수행하기도 한다[6].

최근에는 인공 지능을 적용하여 음성 인식 기술이 연구되고 있다. 머신 러닝을 기반으로 한 DNN(Deep Neural Network) 기술과 음성 인식 기술이 융합되어 자동 음성 인식(ASR) 시스템의 성능이 크게 향상되어 여러 깨끗한 환

경에서 우수한 성능을 보여주기도 한다. 이러한 ASR 시스템은 잡음에 취약하여 음성 신호가 잡음에 심하게 손상되면 성능이 크게 저하될 수 있다. 그러므로 여러 사람이 말을 하는 경우, 다양한 종류의 잡음 환경, SNR을 변화시키는 다양한 시나리오에서 음성 인식 성능을 향상시키는 연구가 진행되고 있다. DNN은 많은 양의 훈련 데이터로 학습함으로써 잡음이 많은 음성 오디오에서 더 깨끗한 신호와 특징을 얻을 수 있어서 다양한 기법의 DNN 기술이 음성 인식 기술에 적용되고 있다[1].

음성인식에 이용되는 인공 지능 기술은 다음과 같이 정리가 될 수 있다. 인공지능을 구성하는 기본 단위인 단일 노드는 생물학적 뉴런과 유사하다. 노드의 값은 일반적으로 입력의 가중치 합과 비선형 활성화 함수로 계산된다. 이론적으로 단일 노드는 수치 해상도가 허용하는 한 엄청난 양의 정보를 나타낼 수 있다. 실제로 DNN은 여러 개의 신경망 층과 각 층은 여러 개의 노드로 구성된다. 결과적으로, 많은 비선형 활성화 함수를 결합할 때 인공지능 네트워크는 입력과 출력 사이의 복잡한 관계를 학습할 수 있다. DNN에 자주 사용되는 일반적인 신경 계층에는 완전 연결 계층을 이용한 다층 퍼셉트론(Multi-Layer Perceptron, MLP), 컨볼루션 계층을 이용한 CNN(Convolutional Neural Network), 순환 계층을 이용한 RNN(Recursive Neural Network), 대상의 특성 벡터를 뽑아내는 Auto-encoder, 및 상호 적대적 훈련을 시키는 GAN(Generative Adversarial Network) 등이 포함된다[7,8,9].

MLP는 일반적으로 특징 벡터(feature vectors)를 다루는 입력층, 입력에 대응하는 출력 값 또는 확률을 나타내는 출력층, 그리고 입력층과 출력층 사이의 히든층으로 이루어져 있으며 이들은 모두 완전히 연결되어 있다. 그러므로 모든 연결을 갖춘 완전 연결 계층으로 어떤 음성의 특징을 추출하기 위해서 사용되며 그 성능이 전통적인 방식에 비해 우수하다[10].

컨볼루션 신경망(CNN)은 입력, 컨볼루션 레이어, pooling layer, fully connected layer로 구성되어 있다. 컨볼루션 레이어는 영상 신호 처리를 위해 개발된 것으로 영상 신호에 필터를 사용해서 지역적인 적당한 특징을 추출한다. Pooling layer는 전체적인 계산양을 줄이기 위해 특징들을 down sampling을 한다, fully connected layer는 MLP와 비슷하게 출력 신호를 예측한다. 2D CNN은 다음 장의 lip reading에 적용이 되며 입력에 대해 작은 크기의 2D 컨볼루션 필터를 사용하고 얼굴 또는 입술 주위의 영상으로 패턴의 로컬 활성화한다. CNN을 음성 신호에만 이용하는 경우는 음향 신호의 시간-주파수 표현을 이미지로 간주할 수 있기 때문에 시간-주파수로 표현하여 응용할 수 있다. 또한 2D 커널을 1D 커널로 수정하여 원시 신호에 직접 적용할 수도 있다. 최근 연구에 따르면 컨볼루션 계층은 원시 신호에서 기본 주파수를 자동으로 학습할 수 있다. 이외 3D CNN은 연속적인 영상의 일부분을 같이 이용하기 위해 3D 필터를 이용하므로 시공간적인 특성을 효율적으로 추출할 수 있다[11].

순환 신경망(RNN)은 순환 계층을 기본으로 사용하고 순환 연결하여 사용한다. 이러한 연결은 결과적으로 네트워크에 이전에 처리된 입력에 액세스할 수 있는 기능을 이용하며 시계열로 연결된 시퀀스 신호를 해석하기에 적당하다. 그러나 RNN은 훈련 시 기울기가 사라지는 문제로 인해 cost function이 의미 없어져서 장기 시간 컨텍스트 정보에 이용할 수 없을 수 있다. 이러한 한계를 극복하기 위해 LSTM(Long Short-Term Memory) 또는 GRU(Gated Recurrent Unit)이 도입되었으며, 이는 음성과 노이즈의 시간에 따라 크게 달라지는 특성에 적용할 수 있다. 그러므로 LSTM-RNN은 긴 시간 범위에서 음성 및 노이즈의 context 정보를 학습할 수 있게 되었다[7,12].

Auto-encoder는 데이터의 압축 분포를 학습하는 데 사용되는 인공지능기술로 인코더와 디코더로 구성됩니다. 인코더는 고차원 공간의 데이터를 저차원 공간으로 변환하는 반면 디코더는 저차원 데이터를 고차원 데이터로 변환한다. Auto-encoder의 디코더는 음성인식을 위한 특징 추출에 사용되는 인코더에 의해 학습된 압축된 표현이므로 학습에만 사용되고 검증을 위해 폐기된다. 이는 비지도 학습 훈련이므로 레이블된 데이터를 요구하지 않는 장점을 가지고 있다. Auto-encoder는 가능한 한 많은 정보를 캡처하려고 노력하지만 압축을 통한 정보를 잃을 수도 있으며 이럴 경우 입력의 작은 부분만 구성하게 되므로 비효율적일 수 있다[13].

GAN은 모델 향상 성능을 더욱 정교화하기 위해 적대적 훈련을 수행한다. 이 훈련 알고리즘은 두 개의 네트워크, 즉 하나의 생성 네트워크(Generator)와 하나의 판별 네트워크(Discriminator)를 계단식 네트워크 구조로 구현한다. 생성 네트워크는 노이즈가 많은 음성을 깨끗한 음성으로 매핑하여 판별 네트워크를 속이려고 하는 반면, 판별 네트워크는 입력이 향상된 음성(False)에서 왔는지 아니면 깨끗한 음성(True)에서 왔는지를 구별하는 것을 목표로 하여 두 네트워크는 미니맥스 게임을 한다. 적대적 훈련 전략은 다른 전통적인 접근법보다 우수한 성능을 보이는 것으로 밝혀졌다[14,15].

최근에는 확률미분방정식(Stochastic Differential Equation, SDE)을 이용하는 확산모델(diffusion model)기법을 이용해서 음성인식을 향상시키고, Self-Attention 기법을 바탕으로 개발된 transformer를 이용하여 자연어 처리와 영상처리를 함으로 음성 인식을 향상시키는 연구되고 있다. 이는 제4장

에서 다룰 다중 모달 음성인식에 적용에 자주 적용이 되고 있다[16,42].

인공지능 기술을 이용하여 음성 신호로부터 음성 신호의 특징을 추출하여 신호 처리함으로써 잡음이 많은 환경에서 음성의 인식률을 상당히 높였다. 이로 인해 음성 인식의 견고성을 향상되었으며 인공 지능의 발전과 더불어 계속 발전해 나가고 있다. 음성 신호를 기반을 하는 자동 음성 인식 기술은 잡음이 많은 환경에서 음성 인식률을 높이는데 상당한 발전을 하였으나, 아직 아직까지 음성신호가 심하게 손상을 입었거나, 신호자체가 없을 경우에 대해서는 그 성능이 충분하지 않으며, 특히 음성신호의 간섭 등에 약한 특징이 있다.

3. 영상을 이용한 음성 인식 기술

음성 인식 기술은 주로 수신된 음성 신호를 기반으로 신호의 특성을 분석하여 음성을 인식하기 때문에 인식율이 환경에 따라 차이가 크며, 잡음이 많은 환경에서 인식률이 낮아서 영상을 비롯한 다양한 다른 보조의 신호를 이용하여 음성 신호를 만들어 내는 연구가 이루어지고 있다. 특히 음성 신호 자체가 아닌 다른 요소(실제 사람의 발성 없이)를 이용하여 음성 정보를 전달하는 SSI(Silent Speech Interface) 기법이 연구되었다. SSI 기술은 청각에 문제가 있거나 사고 후 또는 후두암 치료 후 영구적인 발성 손상을 입은 언어 장애인에게 유용하게 사용될 수 있다[2,3,4,5].

SSI 기술은 기본적으로 비 음성 신호로부터 음성 신호의 특징 변수를 추출하고 이로부터 해당 음성의 특징 변수를 추정하는 방법을 통해 구현된다. SSI에 사용되는 비음성 신호는 가청 주파수 대역의 간섭 신호에 영향을 적게 받으며 음성 신호와 높은 상관 관계가 있어야 하며 높은 수준의 주변 오디오 간섭에 영향을 받지 않아야 한다. 이러한 조건을 만족하는 다양한 유형의 신호와 방식은 GHz 마이크로파[17]에 의한 도플러 주파수 편이, 성대의 초음파 영상(UI)[18], 입의 시각적 모양[19], 음향 도플러 음파 신호[20,21,22], 비청각 마이크로폰(NAM)[23] 또는 근전도(EMG)[24]에 의해 녹음된 신호 등이다. 그 중 입의 시각적 모양을 이용한 Lip Reading기술과 도플러 기반 방법은 다른 방법에 비해 비접촉으로 신호를 얻을 수 있다는 장점이 있다. 근전도 및 비청각 마이크로폰 기반의 방법은 접촉 감지와 관련된 문제로 어려움을 겪고 있다. 도플러 기반 기술은 입의 영역 추적의 부정확성과 같은 이미지 처리 관련 문제가 없다. 초음파 도플러 기반 기법은 발성하고 있는 입주변에 고정 주파수의 초음파 신호를 방사할 때, 반사파에서 관찰되는 도플러 편이를 검출하여 음성을 합성한다. 이때 관찰되는 도플러 편이는 발성과 연관된 근육의 움직임에 따라 특이적으로 나타나게 되며, 음성 신호와 높은 연관성을 갖게 된다. 기존의 초음파 도플러 기반 기법에서는 도플러 편이가 관찰되는 주파수 대역에서 특징 변수를 추출하고 해당 음성 신호의 스펙트럼을 추정하도록 한다.

많은 SSI와 관련한 연구 중 영상을 이용하는 VSR(Video Speech Recognition) 기술은 음성 신호의 품질과 상관없이 영상으로 화자의 말을 인식하기 때문에, 기존 음성 인식 기술 대비 더욱 높은 인식률을 제공한다. 특히 영상 신호를 이용하는 Lip reading 기술은 언어와 얼굴 해부학 모두에 대한 깊은 이해를 필요로 하기 때문에 단순 신호 처리 이외의 복잡한 기술이지만 그 필요성 때문에 오랜 기간 동안 연구되고 있으며, 최근에는 영상 처리 기술의 발전과 인공지능 기술의 발전에 따라 Lip reading기술 역시 비약적인 발전을 하고 있다.

Lip reading기술은 화자의 입 영역에서 획득한 이미지를 사용하는데, 입 모양이 발화되는 음성에 의해 결정된다는 원리를 기반으로 하고 있다. 그러므로 입 모양의 영상 신호를 획득하는 방법이 매우 중요하며, 획득 조건의 차이로 인한 기준 패턴의 불일치에 따라 인식 정확도가 크게 영향을 받는 단점이 있다. 영상신호의 획득의 어려움과 불일치는 Lip reading 분야 뿐만 아니라 PCB의 부품 위치 판단 오류 및 오류 조정 기술에서도 연구가 되고 있다[25]. 특히 조명 조건의 변화는 이러한 불일치의 주요 원인이 된다[26]. 이러한 문제에 대처하기 위해 강도 정규화, 콘트라스트 향상[27], 히스토그램 균등화[28] 및 이미지 디헤이징[29]을 적용할 수 있지만, 불균일한 조명 또는 극도로 밝거나 어두운 조명 하에서 성능이 제한된다. 또한 과도하거나 잘못된 이미지 처리로 인한 왜곡은 인식 정확도를 더욱 저하시킬 수 있다. 적외선 이미지의 사용은 불균일한 조명 조건 [30]과 관련된 문제에 대한 대안적인 해결책이 될 수 있으며, 이미지가 가시광선이 아닌 보이지 않는 IR 조명 하에서 획득되기 때문에 주변 조명의 영향을 받지 않고 어두운 환경에서도 의미 있는 이미지를 얻을 수 있는 장점이 있다.

Lip reading기술은 단순 숫자/알파벳 인식 문제에 적용되기 시작했으며, 이후 격리된 단어 인식에 적용되고 있고 문장 수준 인식기를 구현하기 위해서는 하위 단어 (예: 음소) 기반 인식 이후 합성하는 형태로 연구가 되어 왔다. Lip reading기술은 크게 두가지의 단계로 이루어진다. 입 주위의 운동을 통해 특징 변수 추출하는 front end와 추출된 특징변수를 이용해서 언어적인 내용으로 분류하는 back end이다. 특징변수 추출 기법은 DNN의 인공지능을 이용하기 전에는 특징 추출을 위해 Linear Discriminant

Analysis(LDA), Principal Component Analysis(PCA), Direct Cosine Transformations(DCTs) discrete wavelet transform (DWT), Active Appearance Models(AAMs) 과 같은 알고리즘을 사용하여 입 주변 Region of Interest (ROI)에서 추출한 픽셀 값을 사용했다. 인공지능을 이용하기 시작한 이후에는 CNN, RNN, Autoencoder, GAN, Transformer등이 이용되고 있다[31,32,33]. Autoencoder는 레이블이 지정되지 않는 비지도 학습을 하면서 고차원 공간의 시각적 특징 데이터를 저차원 공간으로 매핑할 수 있는 장점이 있어서 적용되기 시작하였다. 하지만 CNN은 공간적 및 시간적 특징을 학습할 수 있기 때문에 특징변수 추출 관점에서 가장 효율적인 방식으로 알려지면서 Lip reading 기술을 위한 특징변수 추출 방법으로 가장 많이 사용되고 있다[34,35].

분류(Classification) 방법은 추출된 특징들을 가지고 음성으로 분류하는 것으로 인공지능 적용 초기에는 LSTM과 GRU 형태의 RNN이 주로 사용되어 왔으나 최근 몇 년 동안 트랜스포머가 병렬 계산을 더 잘 수행하고 장기 종속성을 학습하며 더 짧은 시간에 훈련할 수 있기 때문에 RNN을 대체하기 시작했다. 그 분류 단위는 음소(sub-word), 단어, 문장 수준으로 진행되고 있다. 단어 수준 단위의 경우, Lip reading이 단순 단어의 분류 작업일 수 있어서 일반적으로 높은 정확도를 달성할 수 있다. 문자 또는 음소 수준을 가진 Lip reading은 언어의 가장 작은 소리 단위인 음소를 인식하는 것으로 전체 단어를 인식하는 전통적인 Lip leading에 비해 상당한 발전을 하였다. 음소 단위의 인식 이후 단어를 조합하기도 한다. 문장 수준에 대한 Lip reading 연구도 진행되고 있는데 방대한 양의 비디오로 구성된 데이터 세트를 이용하여 문장 수준 Lip reading에 대한 연구가 진행되고 있다[5].

Lip reading의 특징변수 추출 단계(front end)과 분류(classification) 단계(back end)의 두개의 각기 다른 인공지능 기법을 적용하거나 여러 개의 인공지능 기법을 혼합하여 사용하여 성능을 향상시키도 하는데, 프론트 엔드에 CNN과 2-D ResNet을 back-end에 BI-LSTM을 사용하여 시간 정보를 캡처하여 상당한 성능 개선을 할 수 있다[36,37]. 이외에 고르지 못한 조명 조건과 관련된 문제들을 해결하기 위해 적외선(IR) 영상을 사용할 수 있고, 이런 영상들을 일반적인 RGB 카메라 센서를 이용해서 비교적 높은 품질의 근적외선(NIR) 이미지를 획득할 수 있다. 참고문헌 [38]에서는 다양한 조명 조건에서 화자의 입술 근처에서 얻은 NIR 데이터를 이용하여 3D DFT와 DNN기술을 적용하여 80개의 단어를 분류하였다.

전반적으로, 영상을 이용하는 Lip reading분야는 빠르게 발전하고 있으며, 특히 난청을 가진 사람들에게 많은 도움을 줄 수 있어 사람들의 삶에 긍정적인 영향을 미치고 있다. 하지만 아직 그 한계가 남아 있는데 훈련되지 않는 단어나 소리에 인식률이 낮으며 훈련에 이용된 데이터에 나타나지 않은 화자에 대해 인식률이 낮아져 일반화하기에 부족한 면이 있다. 즉 화자 의존성 해결, 다양한 공간과 해상도에 적용 등 모두 경우에 이용이 될 수 있도록 일반화해야 할 필요성이 과제가 남아 있다. 또한 입술 또는 얼굴 영상 신호로 충분한 인식률을 얻지 못할 경우 추가 신호를 사용할 수 있다.

4. 음성 신호와 영상 신호를 동시에 이용하는 음성 인식 기술

음성인식 기술은 오디오, 텍스트, 또는 영상 신호 같은 서로 다른 modality를 결합하여 신호처리를 수행하면서 전반적인 성능을 향상시키고 있다[39,40,41]. 시청각 음성 인식 (Audio Video Speech Recognition, AVSR)은 특히 시끄럽거나 어려운 환경에서 음성 인식의 정확도를 향상시키기 위해 영상 신호 처리와 음성 인식을 결합한 다중 modality 처리를 하는 분야로, 입술 움직임, 표정과 같은 시각 정보와 음향 신호와 같은 청각 정보를 모두 활용하여 음성 인식을 수행한다. 특히 인공지능의 딥 러닝은 시청각 음성 인식에 정확도를 크게 향상시켰다. 시각적 특징과 청각적 특징을 효과적으로 추출하고 통합하기 위해 컨볼루션 신경망(CNN) 및 순환 신경망(RNN)과 같은 다양한 딥 러닝 아키텍처를 이용하고 있다. 시각 정보와 청각 정보를 효과적으로 결합하기 위해 고급 다중 modality 융합 기술이 필요하며 서로 간 보완 정보를 캡처하여 보다 강력하고 정확한 음성 인식이 가능하다.

시청각 음성 인식은 더 포괄적이고 지능적인 시스템을 만들기 위해 자연어 처리 (NLP) 및 화자 인식과 같은 다른 기술과 통합되고 있으며, attention based transformer를 이용하면서 단어 뿐만 아니라 단어 뒤에 있는 맥락과 의도도 이해할 수 있도록 한다. Self-Attention을 이용하여 입력과 출력요소 사이에 가중치를 부여하고 가중치가 높은 입력 또는 출력 요소들의 정보를 집중적으로 반영하므로 긴 문장을 처리하는데 유용하게 하고 cross-attention을 이용하여 음성 인식을 위해 가장 관련성이 높은 시각 및 청각 기능에 초점을 맞추도록 한다. 이런 attention based 인공지능 기법을 이용하여 소음이 많거나 어려운 환경을 처리하는 시스템의 성능이 향상되었다[42,43].

새로운 모델 아키텍처와 대규모 데이터 수집에 대한 최근의 발전은 시청각 음성 인식 성능을 새로운 수준으로 끌어올렸다. 그럼에도 불구하고 많은 신경망 아키텍처는 대규모 훈련 데이터를 이용하는 지도 학습에만 초점을 맞추었기 때문에, 기존의 시청각 음성 인식 연구도 완

전한 지도 학습으로 비용이 많이 드는 레이블이 지정된 데이터가 필요로 하였다[44]. 레이블이 지정되지 않은 방대한 데이터를 활용하기 위해 비지도 학습의 필요에 따라 비지도 학습의 한 형태인 Self-Supervised Learning 연구되고 있다. 최근 제안된 AV-Visual HuBERT(AV-HuBERT)는 시청각 음성 인식 연구자들에게 많이 이용되는 워크 프레임인데 자기 지도 학습을 사용한다. 이는 레이블이 지정되지 않은 대량의 시청각 음성 데이터를 사용하여 소리와 관련 입술 움직임 사이의 미묘한 상관 관계를 캡처하도록 모델을 사전 훈련한 다음, 최상의 시청각 음성 인식 성능을 위해 모델을 미세 조정하는 데 소량의 레이블된 시청각 음성 데이터만 사용한다[45].

레이블이 지정되지 않은 데이터를 이용하는 Self-Supervised Learning은 학습을 하기 위해 사용할 데이터의 레이블 및 손실 기능을 결정하는 Pretext tasks 기능에 의해 성능의 차이가 발생한다. 음성 신호에서 Pretext tasks는 화자의 의존도 및 더해지는 잡음의 종류와 크기를 변화시키는 것일 수 있고, 영상 신호에서 Pretext tasks은 회전된 이미지의 각도를 예측하거나, 이미지에서 분할된 영역의 상대적 위치를 학습하거나, 셔플된 패치를 다시 배치하거나, 그레이 스케일 입력 이미지에 색을 더하는 것일 수 있다. 연속된 영상 신호 기반 Pretext tasks은 비디오에서 움직이는 객체를 추적하거나, 시간 프레임 순서를 검증하거나, 비디오 색상화 등일 수 있다[44,45].

Self-Supervised Learning기술은 자연어 처리에 사용되면서 주변 단어를 사용하여 중심 단어를 예측하거나 그 반대의 경우, autoregressive fashion으로 이전 단어를 조건화하여 다음 단어를 생성하고, 마스크 토큰 또는 연속 발화를 완료하고, 셔플된 단어의 순서를 복구하거나 회전된 문장의 순열을 복구하는 pretext 방법도 제안되고 있다. 또한 오디오와 시각적 양식이 의미론적으로 일관되거나 시간적으로 동기화되는 서로 다른 양식 간의 강한 상관 관계를 이용하는 다중 모달 또는 크로스 모달 자기 지도 학습을 방법도 연구되고 있다. 다중 모달 자기 지도 학습은 공동 또는 공유 잠재 공간을 학습하는 것을 목표로 하는 반면, 크로스 모달 자기 지도 학습은 한 양식이 다른 양식을 감독할 수 있도록 한다. 다중 모달의 데이터 사이에는 서로 간의 신호 일치 및 동기화가 필요하다. 즉 영상 신호와 음성 신호 사이의 일치 여부를 예측하거나, 신경망이 소리를 분류하거나, 교차 모드 검색을 학습하거나, 이미지에서 음원을 찾을 수 있도록 해야 한다[46,47,48].

전반적으로 AVSR 분야는 빠르게 발전하고 있으며, 실제 시나리오에서 음성 인식을 향상시킬 수 있는 상당한 잠재력을 가지고 있어서 향후 널리 적용될 것으로 예상된다.

5. 결 론

본 논문에서는 최근 연구되고 있는 음성 인식 기술과 미래 유망 기술들을 종합적으로 살펴보았다. 다양한 인공지능 기술의 적용을 통해 잡음이 많은 환경에서 음성 인식의 견고성을 향상시키고, 자동 음성 인식(ASR) 시스템의 성능을 크게 발전시켰으며, 다중 모달 접근법을 통해 음성 뿐만 아니라 영상 신호를 결합하여 음성의 인식률을 향상시키고 있다. 앞으로의 음성 인식 기술은 대규모 훈련 데이터를 효과적으로 활용하거나, 레이블이 없는 데이터를 활용한 Self-Supervised Learning과 같은 새로운 학습 기법이 도입될 것이고, 이러한 기술들은 음성 인식 시스템을 더욱 견고하고 실용적으로 만들어 다양한 분야에서 적용될 가능성이 크다.

그러나 여전히 해결해야 할 과제가 남아 있다. 현재 음성 인식 시스템은 다양한 잡음 환경에서의 성능 저하 문제를 완전히 극복하지 못하고 있으며, 특정 환경에서의 성능 차이가 존재한다. 특히 SSI기술의 핵심인 Lip reading에 인공 지능 기법을 적용하기 위해서는 영상 신호의 획득이 중요하다. 좋은 품질의 영상신호를 획득하기 위해 조명의 변화, 적외선 촬영 등의 외부적인 품질 향상에 노력을 하지만, 입술 주위의 좋은 영상 신호를 획득하기가 어렵다. 트랜스포머 계열의 대규모 데이터를 사용하는 인공지능 기법에서는 기존 녹화된 영상신호의 디스플레이 화질이 중요하므로 이를 향상시키는 연구가 필요하다. 이외에도 한국어와 같이 어근과 접사가 결합되는 언어에 대한 연구가 부족한 상황이며, 이러한 연구가 강화되어야만 보다 포괄적이고 정확한 음성 인식 시스템이 개발될 수 있을 것이다. 또한, 음성 인식 기술이 일상 생활에서 널리 사용되기 위해서는 성능 향상 뿐만 아니라 사용자 편의성, 소형화, 가격 경쟁력 등 실용적인 측면에서의 발전도 이루어져야 한다.

참고문헌

1. G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kings bury, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82–97, 2012.
2. A. Fernandez-Lopez, F. M. Sukno, “Survey on automatic lip-reading in the era of deep learning,” Image and Vision Computing, vol. 78, pp. 53–72. 2018.
3. J. A. Gonzalez-Lopez, A. Gomez-Alanis, J. M. Martín Doñas, J. L. Pérez-Córdoba and A. M. Gomez, "Silent Speech Interfaces for Speech Restoration: A Review," in

IEEE Access, vol. 8, pp. 177995-178021, 2020.
4. M. Hao, M. Mamut, N. Yadikar, A. Aysa and K. Ubul, "A Survey of Research on Lipreading Technology," in IEEE Access, vol. 8, pp. 204518-204544, 2020.
5. S. Fenghour, D. Chen, K. Guo, B. Li and P. Xiao, "Deep Learning-Based Automated Lip-Reading: A Survey," in IEEE Access, vol. 9, pp. 121184-121205, 2021.
6. K. Paliwal, K. Wójcicki, B. Shannon, "The importance of phase in speech enhancement," Speech Communication, vol. 53, No. 4, pp. 465-494, 2011.
7. M. Wöllmer, B. Schuller, F. Eyben and G. Rigoll, "Combining Long Short-Term Memory and Dynamic Bayesian Networks for Incremental Emotion-Sensitive Artificial Listening," in IEEE Journal of Selected Topics in Signal Processing, vol. 4, no. 5, pp. 867-881, 2010.
8. J. T. Geiger, F. Weninger, J. F. Gemmeke, M. Wöllmer, B. Schuller and G. Rigoll, "Memory-Enhanced Neural Networks and NMF for Robust ASR," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 6, pp. 1037-1046, 2014.
9. Y. Qian, M. Bi, T. Tan, and K. Yu, "Very deep convolutional neural networks for noise robust speech recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 12, pp. 2263-2276, 2016.
10. G. E. Dahl, D. Yu, L. Deng and A. Acero, "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 1, pp. 30-42, Jan. 2012.
11. G. Trigeorgis et al., "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, pp. 5200-5204, 2016.
12. H. Joo, K. Lee, "Estimating speech parameters for ultrasonic Doppler signal using LSTM recurrent neural networks," The Journal of the Acoustical Society of Korea, vol.38, no.4, pp. 433-441, 2019
13. Z. Zhang, N. Cummins and B. Schuller, "Advanced Data Exploitation in Speech Analysis: An overview," in IEEE Signal Processing Magazine, vol. 34, no. 4, pp. 107-129, July 2017.
14. K. Lee, "An acoustic Doppler-based silent speech interface technology using generative adversarial networks'" The Journal of the Acoustical Society of Korea. vol.40, no.2, pp. 161-168, 2021.
15. A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," IEEE Signal Processing Magazine, vol. 35, pp. 53-65, 2018.
16. Julius Richter, Simon Welker, J-M Lemercier, Bunlong Lay, Timo Gerkmann, "Speech Enhancement and Dereverberation with Diffusion-Based Generative Models," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 2351-2364. 2023.
17. C. -S. Lin, S. -F. Chang, C. -C. Chang and C. -C. Lin, "Microwave Human Vocal Vibration Signal Detection Based on Doppler Radar Technology," in IEEE Transactions on Microwave Theory and Techniques, vol. 58, no. 8, pp. 2299-2306, Aug. 2010.
18. B. Denby, T. Schultz, K. Honda, T. Hueber, J.M. Gilbert, J.S. Brumberg, "Silent speech interfaces," Speech Communication, vol. 52, No. 4, pp. 270-287, 2010.
19. T. Le Cornu and B. Milner, "Generating Intelligible Audio Speech from Visual Speech," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 9, pp. 1751-1761, 2017.
20. K. Lee, "Ultrasonic Doppler Based Silent Speech Interface Using Perceptual Distance," Applied Sciences. 12(2), 827, 2022.
21. K. Lee, "Speech enhancement using ultrasonic doppler sonar", Speech Communication, Vol. 110, pp. 21-32, July 2019.
22. K. Lee, "Silent Speech Interface Using Ultrasonic Doppler Sonar," EICE Transactions on Information and Systems, vol. E103.D, no. 8, pp. 1875-1887, 2020.
23. T. Toda, M. Nakagiri and K. Shikano, "Statistical Voice Conversion Techniques for Body-Conducted Unvoiced Speech Enhancement," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 9, pp. 2505-2517, Nov. 2012.
24. M. Janke and L. Diener, "EMG-to-Speech: Direct Generation of Speech From Facial Electromyographic Signals," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 12, pp. 2375-2385, Dec. 2017.
25. G. Shin, J. Kim, "A Study on the Intelligent Recognition of a Various Electronic Components and Alignment Method with Vision," Journal of the Semiconductor & Display Technology, vol. 23, no. 2, pp. 1-5, 2024.
26. X. Tan and B. Triggs, "Enhanced Local Texture Feature Sets for Face Recognition Under Difficult Lighting Conditions," in IEEE Transactions on Image Processing, vol. 19, no. 6, pp. 1635-1650, 2010.
27. Á. Chavarín, E. Cuevas, O. Avalos, J. Gálvez and M. Pérez-Cisneros, "Contrast Enhancement in Images by Homomorphic Filtering and Cluster-Chaotic Optimization," in IEEE Access, vol. 11, pp. 73803-73822, 2023.
28. P. -H. Lee, S. -W. Wu and Y. -P. Hung, "Illumination Compensation Using Oriented Local Histogram Equalization and its Application to Face Recognition," in IEEE Transactions on Image Processing, vol. 21, no. 9, pp. 4280-4289, Sept. 2012.
29. M. Zheng, G. Qi, Z. Zhu, Y. Li, H. Wei and Y. Liu, "Image Dehazing by an Artificial Image Fusion Method

Based on Adaptive Structure Decomposition," in IEEE Sensors Journal, vol. 20, no. 14, pp. 8062-8072, 15 July15, 2020.
30. D. Sugimura, T. Mikami, H. Yamashita and T. Hamamoto, "Enhancing Color Images of Extremely Low Light Scenes Based on RGB/NIR Images Acquisition With Different Exposure Times," in IEEE Transactions on Image Processing, vol. 24, no. 11, pp. 3586-3597, Nov. 2015.
31. Y. Kumar, R. Jain, K. M. Salik, R. R. Shah, Y. Yin, R. Zimmermann, “Lipper: Synthesizing thy speech using multi-view lipreading,” in Proc. AAAI Conf. Artif. Intell., vol. 33, pp. 2588–2595, 2019.
32. K. Vougioukas, P. Ma, S. Petridis, and M. Pantic, “Video-driven speech reconstruction using generative adversarial networks,” in Proc. Interspeech, Sep. 2019, pp. 4125–4129.
33. M. Wand, J. Koutník, and J. Schmidhuber, “Lipreading with Long Short-Term Memory,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. 2016, pp. 6115–6119.
34. A. Ephrat and S. Peleg, “Vid2Speech: Speech reconstruction from silent video,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. 2017, pp. 5095–5099.
35. H. Akbari, H. Arora, L. Cao, and N. Mesgarani, “Lip2Audspec: Speech reconstruction from silent lip movements video,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Apr. 2018, pp. 2516–2520.
36. T. Stafylakis and G. Tzimiropoulos, “Combining residual networks with LSTMs for lipreading,” in Proc. Interspeech, Aug. 2017, pp. 3652–3656.
37. B. Martinez, P. Ma, S. Petridis, and M. Pantic, “Lipreading using temporal convolutional networks,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2020, pp. 6319–6323.
38. K. Lee, “Improving the Performance of Automatic Lip-Reading Using Image Conversion Techniques,” Electronics, 13(6), 1032, March 2024.
39. M. Sadeghi, S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, “Audio-visual Speech Enhancement Using Conditional Variational Auto-Encoders,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 1788-1800. 2020.
40. X. Qian, Z. Wang, J. Wang, G. Guan, and H. Li, “Audio-visual cross-attention networks for robotic speaker tracking,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 550-562. 2023.
41. Lei. Liu, Li Liu, and H. Li, “Computation and Parameter Efficient Multi-Modal fusion Transformer for Cued Speech Recognition,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 32, pp. 1559-1572. 2024.
42. B. Shi, W. Hsu, A. Mohamed, “Robust Self-Supervised Audio-Visual Speech Recognition,” arXiv:2201.01763, 2022.
43. L. Qu, C. Weber, S. Wermter, “LipSound2: Self-Supervised Pre-Training for Lip-to-Speech Reconstruction and Lip Reading,” IEEE Transactions on Neural Networks and Learning systems, vol. 35, no. 2, pp. 2772-2782, 2024.
44. T. Afouras, J. Chung, A. Senior, O. Vinyals, A. Zisserman, "Deep Audio-Visual Speech Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 12, pp. 8717-8727, 2022.
45. C. Xie, T. Toda, “Noisy-to-Noisy Voice Conversion Under Variations of Noisy Condition,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 3871-3882. 2023.
46. J. Devlin, M. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 4171–4186, 2019.
47. Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “ALBERT: A lite BERT for self-supervised learning of language representations,” ICLR 2020 Conference, 2019, arXiv:1909.11942.
48. B. Chen et al., "Multimodal Clustering Networks for Self-supervised Learning from Unlabeled Videos," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 7992-8001.

접수일: 2024년 9월 10일, 심사일: 2024년 9월 13일,
게재확정일: 2024년 9월 14일