논문 2025-4-19 http://dx.doi.org/10.29056/jsav.2025.12.19

디지털 취약 계층의 디지털 사이니지 사용성 향상을 위한 음성-영상 기반 UI 파이프라인 설계

김정현*, 박종섭*, 최용수**†

Designing a Voice-Video Based UI Pipeline to Improve Digital Signage Usability for Digitally Vulnerable Groups

JungHyun Kim*, JongSub Park*, YongSoo Choi**†

요 약

본 논문은 교통 약자의 디지털 사이니지 사용성 향상을 위한 음성-영상 융합형 UI 파이프라인을 제안하며 크게 시각적 발화 감지와 청각적 신호 개선의 두 단계로 구성된다. 첫째, 영상 처리 모듈은 Google MediaPipe Holistic과 MobileNetV2-GRU 하이브리드 모델을 활용하여 사용자의 입술 움직임을 분석하고 실시간으로 '발화 의도'를 포착한다. 특히 데이터 증강(Data Augmentation)과 순환 버퍼(Circular Buffer)를 통해 초기 발화 손실을 방지하고 100%의 발화 감지 정확도를 달성하였다. 둘째, 음성 처리 모듈은 '불필요한 개입 최소화(Do No Harm)' 원칙에 기반한 적응형 SNR(Signal-to-Noise Ratio) 노이즈 제거 알고리즘을 적용한다. 딥러닝 모델(Sepformer)이 한국어 음성 신호를 왜곡하는 문제를 해결하기 위해, SNR 6dB를 임계값으로 설정하여 저SNR 환경에서는 선별적 노이즈 제거를 수행하고, 고SNR 환경에서는 처리를 생략(Skip)하여 음성 손실을 방지한다. 특히, 저SNR환경에서 전체 데이터의 79.4% 음성처리를 생략할 경우, WER 1.07% 감소시키는 효과를 보였다. 또한 시각적 발화 감지에서 데이터 증강 기법을 적용하여 정확도 1.0(100%), 검증 손실(Loss) 0.0004을 획기적으로 달성할 수 있다.

Abstract

This paper proposes a speech-image UI pipeline to improve the usability of digital signage for the mobility-challenged. It consists of two main stages: visual speech detection and auditory signal enhancement. First, the image processing module utilizes Google MediaPipe Holistic and a MobileNetV2-GRU hybrid model to analyze the user's lip movements and capture "speech intent" in real time. Specifically, data augmentation and a circular buffer prevent early speech loss, achieving 100% speech detection accuracy. Second, the speech processing module adopts an adaptive signal-to-noise ratio (SNR) noise removal algorithm based on the "Do No Harm" principle. To solve the problem of deep learning models (Sepformer) distorting Korean speech signals, the SNR threshold is set to 6 dB. It is remove noise in case of low-SNR environments and skip processing in case of high-SNR environments to prevent speech loss. In particular, if it omit the speech processing of 79.4% of the total speech data in a low-SNR environment, 1.07% decrease in WER will be achieved. Furthermore, applying data augmentation techniques in visual speech detection significantly achieve the accuracy of 1.0(100%) and the loss of 0.0004.

한글키워드 : 디지털 사이니지, 음성-영상, 잡음 제거, UI 파이프라인, WER

keywords : Digital Signage, Speech-Image, Noise Cancellation, UI Pipeline, WER

* 신한대학교 소프트웨어융합학과
** 신한대학교 미래자동차공학과
† 교신저자: 최용수(email: ciechoi@shinhan.ac.kr)

접수일자: 2025.12.05. 심사완료: 2025.12.13.
게재확정: 2025.12.20.

1. 서 론

디지털 기술의 발전으로 공공장소의 정보 전달 체계가 단방향 정보 전달 시스템에서 양방향 상호작용 체제로 변하고 있다. 하지만 이러한 변화는 장애인, 고령자 등 디지털 정보 소외 계층에게 새로운 장벽을 만들고 있다. 실제로 과학기술정보통신부와 한국지능정보사회진흥원의 「2024 디지털정보격차 실태조사」에 따르면, 일반 국민의 디지털 정보화 수준을 100%로 볼 때 장애인은 83.5%, 고령층은 71.4% 수준에 그쳐 이들 계층이 겪는 정보 격차가 명확히 확인된다[1].

한국소비자원의 조사에 따르면 키오스크 이용에 있어 인지적·심리적 장벽이 매우 높으며 물리적·소통 장벽 역시 심각하다. 수도권의 키오스크 20대를 점검한 결과 85.0%(17대)가 휠체어 사용자의 조작이 어려운 높이(1,220mm 초과)에 설치되어 있었고, 조사 대상 100%(20대)가 시각장애인을 위한 음성 안내나 점자 표시가 전무했다. 또한 주차장 키오스크(5대, 25.0%)의 경우 유일한 소통 수단이 '전화 통화' 방식의 '호출' 버튼뿐이어서 청각장애인의 접근성을 원천적으로 배제하고 있었다[2].

터치 기반 및 고정형 인터페이스가 야기하는 물리적, 인지적, 심리적 장벽을 근본적으로 해결하기 위해서는, 사용자의 조작 없이도 의도를 파악하고 소음이 많은 공공 환경에서도 명확하게 소통할 수 있는 새로운 방식의 인터페이스가 절실히 요구된다. 본 연구에서는 이러한 한계를 극복하고, 음성-영상 기반 UI 파이프라인을 통해 화자 탐지 및 적응형 잡음 제거 시스템을 구현하고자 한다. 이를 위해 시각 정보 기반의 발화 감지와 한국어 음성 특성을 보존하는 SNR 기반 가변적 노이즈 제거 알고리즘을 설계 및 제안한다.

2. 관련 연구 및 연구 동기

2.1 키오스크 사용성 연구

디지털 키오스크의 보급이 확대됨에 따라 장애인 및 고령자의 사용성 문제가 주요 연구 주제로 부상하고 있다. Msweli 등[3]은 시각장애인을 위한 키오스크 인터페이스 설계 가이드라인을 제시하며, 음성 피드백과 촉각 인터페이스의 중요성을 강조하였다. Brunner 등[4]은 고령자의 키오스크 사용 패턴을 분석하여 인지 부하를 줄이는 단순화된 UI 설계 원칙을 도출하였다. 국내에서는 한국정보화진흥원이 「무인정보단말기 UI 플랫폼」 가이드라인을 발표하여 휠체어 사용자를 위한 화면 높이 조절, 시각장애인을 위한 음성 안내 등의 기준을 제시하였다[5]. 그러나 기존 연구들은 주로 물리적 인터페이스 개선에 초점을 맞추고 있으며, 소음 환경에서의 음성 인식 품질 향상에 대한 연구는 상대적으로 부족한 실정이다.

본 연구는 이러한 한계를 보완하여 공공장소의 소음 환경에서도 안정적으로 작동하는 음성 기반 인터페이스를 제안한다.

2.2 시각 기반 발화 감지 기술

시각 정보를 활용한 발화 감지(Visual Speech Detection)는 오디오 신호만으로는 어려운 화자 식별 및 발화 구간 탐지 문제를 해결하기 위해 연구되어 왔다. Tao 등[6]이 제안한 TalkNet-ASD는 오디오-비주얼 특징을 융합하여 다중 화자 환경에서 실제 발화자를 탐지하는 Active Speaker Detection 모델로, 자기지도학습(Self-supervised Learning) 기반의 사전학습을 통해 높은 정확도를 달성하였다. Chung 등[7]은 입술 움직임만으로 발화 내용을 인식하는 Lip Reading 연구를 수행하여 LRS(Lip Reading Sentences) 데이터셋과 함께 딥러닝 기반 모델을

제시하였다. 최근에는 MediaPipe Face Mesh와 같은 경량화된 얼굴 랜드마크 추출 기술이 등장하여 엣지 디바이스에서도 실시간 얼굴 분석이 가능해졌다[8].

본 연구에서는 MediaPipe의 경량성과 MobileNetV2-GRU의 시공간 분석 능력을 결합하여 키오스크 환경에 최적화된 발화 감지 모델을 설계한다.

2.3 딥러닝 기반 음성 향상 기술

음성 향상(Speech Enhancement) 및 음원 분리(Source Separation) 분야에서는 딥러닝 기반 모델이 괄목할 만한 성과를 거두고 있다. Subakan 등[9]이 제안한 Sepformer는 Transformer 아키텍처를 음원 분리에 적용하여 WSJ0-2mix 벤치마크에서 최고 수준의 SI-SDRi(Scale-Invariant Signal-to-Distortion Ratio improvement)를 달성하였다. Luo 등[10]의 Conv-TasNet은 시간 영역에서 직접 음원 분리를 수행하는 경량화된 모델로, 실시간 처리가 가능하다는 장점이 있다. Defossez 등[12]의 Demucs는 U-Net 구조를 활용하여 음악 및 음성 분리에서 높은 성능을 보였다. 그러나 이러한 모델들은 대부분 영어 음성 데이터셋(WSJ0-2mix, LibriMix, VCTK 등)으로 학습되어 있어, 한국어와 같이 음소 체계가 상이한 언어에 적용 시 성능 저하가 발생할 수 있다. 특히 한국어의 된소리(ㄲ, ㄸ, ㅃ, ㅆ, ㅉ), 격음(ㅋ, ㅌ, ㅍ, ㅊ), 치찰음(ㅅ, ㅆ, ㅈ) 등은 영어에 존재하지 않는 음소로, 영어 데이터로 학습된 모델이 이를 노이즈로 오인할 가능성이 높다.

본 연구에서는 이러한 언어적 특성을 고려하여 딥러닝 모델의 무분별한 적용 대신 조건부 처리 전략을 채택한다.

2.4 음성 인식 모델

자동 음성 인식(Automatic Speech Recognition, ASR) 분야에서는 대규모 데이터 학습을 통한 범용 모델이 주목받고 있다. Radford 등[11]이 개발한 Whisper는 68만 시간의 다국어 음성 데이터로 학습된 대규모 ASR 모델로, 별도의 미세조정(Fine-tuning) 없이도 다양한 언어와 도메인에서 강건한 성능을 보인다. 특히 Whisper는 학습 과정에서 다양한 소음 조건의 데이터를 포함하여 내장된 잡음 내성(Built-in Noise Robustness)을 확보하였다. Baevski 등[13]의 Wav2Vec 2.0은 자기지도학습 기반의 사전학습을 통해 적은 양의 레이블 데이터로도 높은 인식률을 달성하였으며, Gulati 등[14]의 Conformer는 CNN과 Transformer를 결합하여 지역적 특징과 전역적 문맥을 동시에 포착한다.

본 연구에서는 한국어 인식 정확도와 잡음 내성이 우수한 Whisper large-v3를 최종 STT 모델로 채택하였으며, 이 모델의 내재적 잡음 내성을 최대한 활용하는 방향으로 전처리 파이프라인을 설계한다.

2.5 SNR 기반 적응형 처리

신호 대 잡음비(SNR)를 기준으로 처리 방식을 분기하는 적응형 접근법은 통신 및 오디오 처리 분야에서 널리 활용되어 왔다. Loizou[15]는 저서에서 SNR 추정 기법과 이를 활용한 적응형 노이즈 제거 알고리즘을 체계적으로 정리하였다. 최근에는 딥러닝 모델의 연산 비용 문제를 해결하기 위해 입력 신호의 특성에 따라 처리 여부를 결정하는 조건부 연산(Conditional Computation) 연구가 활발히 진행되고 있다[16]. 그러나 기존 연구들은 주로 통신 품질 개선이나 연산 효율성에 초점을 맞추고 있으며, 후단 ASR 모델의 성능까지 고려한 종단 간(End-to-End) 최적화 연구는 부족하다.

본 연구는 Whisper 모델의 잡음 내성 특성을

분석하고, 이를 기반으로 SNR 6dB를 임계값으로 설정하여 불필요한 전처리를 생략하는 'Do No Harm' 전략을 제안한다.

3. 제안하는 음성-영상 기반 UI 파이프라인

3.1 Phase 1: MediaPipe Holistic 기반 사용자 탐지 및 ROI 추출

안정적인 상호작용을 위해서는 키오스크 앞에 위치한 주 사용자를 신속하게 탐지하고, 얼굴 영역을 정확히 추출하는 것이 필수적이다. 본 시스템은 기존의 무거운 객체 탐지 모델(Object Detection) 대신, 경량화된 Google MediaPipe Holistic 솔루션을 채택하여 인체 자세(Pose)와 얼굴 랜드마크(Face Mesh)를 단일 파이프라인으로 동시에 처리한다.

3.1.1 사용자 접근 감지 (Distance Estimation)

별도의 깊이 카메라(Depth Camera) 없이 단안 카메라(RGB)만으로 사용자의 접근을 판단하기 위해, MediaPipe Pose 랜드마크 중 좌측 어깨(인덱스 11)와 우측 어깨(인덱스 12) 사이의 유클리드 거리(Euclidean Distance)를 계산한다. 이 어깨 너비 값이 사전 설정된 임계값(SHOULDER_WIDTH_THRESHOLD)을 초과할 경우, 사용자가 키오스크와 상호작용 가능한 거리 내에 진입한 것으로 간주하여 시스템을 활성화한다.

3.1.2 순환 버퍼(Circular Buffer)를 이용한 초기 발화 손실 방지

일반적인 VAD 시스템은 발화가 감지된 시점부터 녹음을 시작하기 때문에, 첫 음절이나 단어가 잘리는 '초기 발화 손실(Front-end Clipping)'

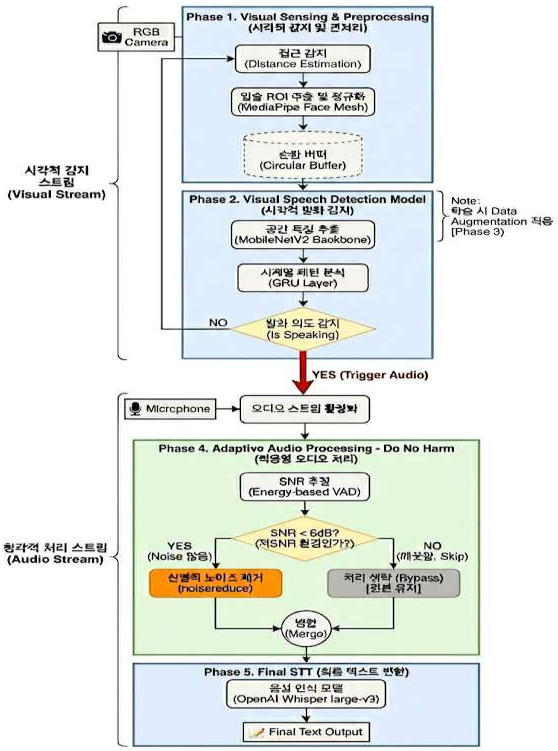

그림 1. 음성-영상 기반 UI 파이프라인
Fig. 1. Speech-Image Based UI Pipeline

문제가 발생한다. 이를 해결하기 위해 본 시스템은 순환 버퍼(Circular Buffer) 메커니즘을 도입하였다. 시스템은 L=90 (약 3초 분량)의 고정 길이를 가진 FIFO(First-In-First-Out) 큐(Queue)를 유지한다. 카메라는 항상 활성화되어 있으며, 매 프레임마다 전처리된 입술 이미지가 큐에 삽입되고 가장 오래된 프레임은 삭제된다. 발화 의도가 감지되는 즉시 큐에 저장된 과거 3초간의 데이터를 모델의 입력으로 사용함으로써, 발화 시작 전의 미세한 입술 움직임까지 놓치지 않고 분석할 수 있다.

3.1.3 정밀 입술 ROI 추출 및 정규화

발화 감지의 정확도를 높이기 위해 전체 얼굴이 아닌 입술 영역(Region of Interest, ROI)만을

정밀하게 추출한다. MediaPipe Face Mesh가 제공하는 468개의 랜드마크 중 입술의 외곽선을 구성하는 27개의 핵심 인덱스(0, 17, 37, 39, 40, 61, 84, 91, 146, 181, 185, 267, 269, 270, 291, 314, 321, 375, 405, 409 등)를 활용한다.추출 과정은 다음과 같다. 첫째, 입술 랜드마크들의 x, y 좌표 집합에서 최소값과 최대값을 산출하여 바운딩 박스를 생성한다. 둘째, 입술 움직임에 따른 동적 변화를 수용하기 위해 상하좌우에 10픽셀의 여유 공간(Padding)을 부여한다. 셋째, 추출된 ROI 이미지는 입술의 가로가 긴 형태적 특성을 반영하여 100 X 50 크기로 리사이징(Resizing)한다. 마지막으로, 조명 변화에 강인하도록 픽셀 값을 [0~255] 범위에서 [0.0~1.0] 범위로 정규화(Normalization)하여 모델의 입력으로 전달한다.

3.2 Phase 2: 시공간 특징 기반 발화 의도 감지

추출된 입술 ROI 시퀀스를 분석하여 실제 '발화(Speaking)'와 단순한 입술 움직임(Non-speaking)을 구분하기 위해 2D CNN과 RNN이 결합된 하이브리드 딥러닝 모델을 설계하였다.

3.2.1 MobileNetV2 기반 공간 특징 추출 (Spatial Feature Extraction)

각 프레임(이미지)에서 입술의 형태적 특징을 추출하기 위해 백본(Backbone) 네트워크로 MobileNetV2를 사용하였다. MobileNetV2는 일반적인 합성곱(Standard Convolution) 대신 깊이별 분리 합성곱(Depthwise Separable Convolution)을 사용하여 연산량을 획기적으로 줄인 경량 모델이다. 깊이별 분리 합성곱은 채널별로 필터를 적용하는 Depthwise Convolution과 채널을 혼합하는 Pointwise Convolution(1 x 1)으로 나뉘어 수행되며, 이는 기존 합성곱 대비 연산량을 약 8~9배 감소시키는 효과가 있다. 또한, 역잔차 구조(Inverted Residual Structure)와 선형 병목(Linear Bottleneck)을 도입하여 정보 손실을 최소화하면서도 메모리 사용량을 최적화하였다. 본 시스템에서는 ImageNet으로 사전 학습된 가중치를 사용하여 전이 학습(Transfer Learning)을 수행하며, 각 프레임(100 x 50 x 3)은 MobileNetV2를 통과하여 고차원의 공간 특징 벡터(Spatial Feature Vector)로 변환된다.

3.2.2 GRU 기반 시간적 패턴 분석 (Temporal Analysis)

단일 프레임만으로는 사용자가 말을 하는 중인지, 단순히 입을 벌리고 있는 것인지 구분하기 어렵다. 따라서 연속적인 프레임의 시간적 흐름을 분석하기 위해 GRU(Gated Recurrent Unit) 레이어를 적용하였다. GRU는 LSTM(Long Short-Term Memory)의 변형된 구조로, 출력 게이트(Output Gate)를 제거하고 업데이트 게이트(Update Gate)와 리셋 게이트(Reset Gate)만으로 구성되어 있다. 이는 LSTM 대비 파라미터 수가 적어 학습 속도가 빠르고 연산 비용이 낮으면서도, 시계열 데이터의 장기 의존성(Long-term Dependency)을 학습하는 성능은 대등하다. 본 모델은 90개의 연속된 특징 벡터 시퀀스를 GRU 레이어에 입력하여 시간적 문맥을 분석하고, 완전 연결층(Dense Layer)과 시그모이드(Sigmoid) 활성화 함수를 통해 최종적으로 발화 확률(0~1)을 출력한다.

3.3 Phase 3: 데이터 증강을 통한 강인성 확보

실제 키오스크 환경에서는 사용자의 고개 각도나 위치가 다양하게 나타날 수 있다. 제한된 데이터셋 환경에서 모델의 일반화 성능을 높이고 과적합(Overfitting)을 방지하기 위해, 학습 단계에서 모든 비디오 시퀀스에 대해 좌우 반전(Horizontal Flip) 증강을 수행하였다. 이를 통해

데이터셋의 규모를 2배로 확장하고, 입술 움직임의 대칭적 특성을 모델이 학습하게 함으로써 100%의 검증 정확도를 달성하였다.

3.4 Phase 4: 적응형 SNR 기반 오디오 전처리 (Adaptive Audio Preprocessing)

시각적 모듈이 발화를 감지하면, 수집된 오디오 신호에 대해 적응형 노이즈 제거(Adaptive Noise Reduction)를 수행한다. 본 연구의 핵심 기여는 기존 딥러닝 기반 음원 분리 모델의 한계를 규명하고, 이를 극복하기 위한 조건부 처리 전략을 제시한 점에 있다.

3.4.1 기존 딥러닝 모델의 한계 분석

초기 실험에서 최신 음원 분리 모델인 Sepformer[3]를 적용하였으나, 한국어 음성 인식 성능이 오히려 17.46%p 악화되는 심각한 문제가 발생하였다. 이는 Sepformer가 영어 음성 데이터셋(WSJ0-2mix, LibriMix 등)으로 학습되어, 한국어 고유의 음소 특성(된소리, 격음, 경음 등)을 노이즈로 오인하여 과도하게 억제(over-suppress)하기 때문으로 분석된다. 특히 한국어의 고주파 성분이 집중된 치찰음(/ㅅ/, /ㅆ/, /ㅈ/ 등)과 파찰음이 심각하게 왜곡되어 Whisper 모델이 인식하지 못하는 형태로 변환되었다. 이러한 문제를 해결하기 위해 3.4.2절에서 제안하는 'Do No Harm' 원칙을 수립하였으며, Sepformer 대신 조건부 처리 전략을 채택하였다.

3.4.2 '불필요한 개입 최소화(Do No Harm)' 원칙

딥러닝 모델의 한계로 인해 '불필요한 개입 최소화(Do No Harm)' 원칙을 수립하였다. 핵심 아이디어는 Whisper 모델이 이미 대규모 다국어 데이터(68만 시간)로 학습되어 상당한 수준의 내장된 잡음 내성(built-in noise robustness)을 보유하고 있다는 점에 착안한 것이다. 따라서 노이즈 제거가 반드시 필요한 극한 소음 환경에서만 선별적으로 개입하고, 그 외의 경우에는 원본 음성을 보존하여 불필요한 왜곡을 방지하도록 하였다.

3.4.3 SNR 임계값 선정 (6dB)

최적의 SNR 임계값을 도출하기 위해 4dB, 6dB, 8dB, 10dB의 네 가지 후보에 대해 실험을 수행하였다. 최적의 SNR 임계값을 도출하기 위해 4dB, 6dB, 8dB, 10dB의 네 가지 후보에 대해 실험을 설계하였다. 각 임계값에서의 처리 효율성과 WER 변화를 비교 분석하여 최적값을 선정하였으며, 상세 결과는 5장에서 기술한다.

3.4.4 조건부 처리 알고리즘

입력 오디오의 SNR을 실시간으로 추정한 후, 청크의 SNR 크기에 따라 다음과 같이 분기 처리한다.

만약, 고SNR 환경(SNR $\geq$ 6dB)이라면 Whisper모델의 자체 잡음내성이 충분해 인식이 가능하므로 기존의 오디오를 유지한다. 반대로 저SNR 환경(SNR $<$ 6dB)이라면 다단계 노이즈 제거를 위해 NoiseReduce라이브러리 기반 스펙트럼 차감 알고리즘을 적용(prop_decrease 파라미터=0.7[노이즈 감쇄율 70%])한다. 이로인해 음성 특성을 보존한채 배경 잡음만 설별 감쇄되는 효과를 가지게 될 것이다.

3.4.5 조건부 처리를 위한 오디오 청크 SNR 추정

입력 오디오 청크(30초 단위)에 대해 조건부 처리의 적용을 위해 다음과 같이 식1(SNR)을 추정한다.

$$SNR(dB) = 10 \times log_{10}^{\frac{S}{N}} \quad \text{-----} \quad (1)$$

단, 에너지 기반 VAD(Voice Activity Detection)로 음성 구간과 묵음 구간 분리 음성

구간의 평균 에너지를 신호(S), 묵음 구간의 평균 에너지를 노이즈(N)로 정의한다.

3.5 Phase 5: 최종 텍스트 변환 (STT)

적응형 전처리를 거친 오디오 스트림은 OpenAI Whisper large-v3 모델을 통해 텍스트로 변환된다. Whisper는 68만 시간의 다국어 음성 데이터로 학습되어 기본적으로 높은 잡음 내성을 보유하고 있으며, 앞서 3.4절에서 제안한 적응형 SNR 기반 전처리는 6dB를 임계값으로 설정하여(3.4.3절), 저SNR 환경(SNR < 6dB)에서는 noisereduce 기반 노이즈 제거를 적용하고, 고SNR 환경(SNR ≥ 6dB)에서는 처리를 생략하여 원본 음성을 보존한다(3.4.4절). 이를 통해 Whisper 모델이 다양한 소음 환경에서 안정적으로 작동할 수 있는 조건을 제공한다.

Whisper 모델 선정 시 large-v3를 채택한 이유는 한국어 인식 정확도가 가장 높으면서도, 실시간 처리가 가능한 수준의 연산량을 유지하기 때문이다. 처리 단위는 30초 청크로 설정하여 메모리 효율성과 컨텍스트 유지의 균형을 맞추었다.

4. 실험 환경

4.1 데이터 셋

본 연구는 제안 모델의 성능 검증을 위해 키오스크 환경을 모사한 자체 데이터셋을 구축하였다. 데이터는 '발화(Speaking)'와 '비발화(Non-speaking)'의 두 가지 클래스로 구성된다. 초기 데이터는 클래스당 24개 시퀀스로 수집되었으나, 데이터 부족에 따른 과적합(Overfitting) 방지를 위해 영상의 좌우를 반전시키는 수평 뒤집기(Horizontal Flip)를 적용하였다. 이를 통해 총 96개 시퀀스로 데이터를 확장함과 동시에 얼굴 방향 변화에 대한 모델의 강건성(Robustness)을 확보하였다. 각 시퀀스는 90프레임(약 3초) 길이로 정규화하였으며, 입력 이미지는 100x50 픽셀로 조정하여 MobileNetV2에 입력하였다. 본 데이터셋(96개 시퀀스)은 대규모 학습에 비해 제한적인 규모이나, 다음과 같은 이유로 유효한 학습이 가능하였다. 첫째, ImageNet으로 사전 학습(Pre-trained)된 MobileNetV2 백본을 활용하여 최상위 레이어만 미세조정(Fine-tuning)하는 전이학습(Transfer Learning) 방식을 채택하였다. 이는 소규모 데이터 환경에서도 높은 일반화 성능을 보장한다[17]. 둘째, 수행하는 태스크가 비교적 단순한 이진 분류이므로 적은 데이터로도 학습 효율을 확보할 수 있다. 한편, 오디오 인식 성능 평가는 6가지 소음 환경(Traffic, Construction, Factory, Facility, Outdoor, Complex)의 유효한 정답 텍스트가 존재하는 67개 청크 데이터를 별도로 구성하여 분석하였다.

4.2 영상 실험 환경

제안 모델은 시공간 특징 추출을 위해 MobileNetV2와 GRU(Gated Recurrent Unit)를 결합한 하이브리드 구조를 갖는다. 학습은 NVIDIA RTX 3060 GPU 환경에서 수행되었으며, 전체 데이터의 80%는 학습(Train)에, 20%는 검증(Validation)에 사용하였다. 최적화 도구는 Adam 옵티마이저를, 손실 함수는 이진 교차 엔트로피(Binary Cross-entropy)를 적용하였다.

4.3 음성 실험 환경

음성 인식 성능 평가를 위해 6개의 대표적인 소음 환경에서 데이터를 수집하였다(표 1).

총 68개의 오디오 청크(각 30초)를 수집하였으며, 유효한 GT가 존재하는 67개 청크에 대해 WER(Word Error Rate)을 평가 지표로 채택하였다.

표 1. 6가지 소음환경 데이터 셋
Table 1. 6 Noise environment data set

소음구분	청크 평균 SNR	청크 수
Traffic (교통 소음)	4.7dB	12
Complex (복합 소음)	6.1dB	9
Construction (공사 소음)	16.8dB	11
Factory (공장 소음)	18.8dB	12
Facility (시설 소음)	18.3dB	12
Outdoor (야외 소음)	8.3dB	11

$$WER(\%) = - \ (S + D + I) \ / \ N \times 100 \qquad (2)$$

단, S는 대체 오류, D는 삭제 오류, I는 삽입 오류, N은 정답 텍스트의 총 단어 수를 의미한다.
(S)정답 단어를 다른 단어로 잘못 인식한 횟수
(D)정답에 있는 단어를 인식하지 못하고 누락한 횟수
(I)정답에 없는 단어를 추가로 잘못 인식한 횟수
(N)정답 텍스트(Ground Truth)의 전체 단어 개수

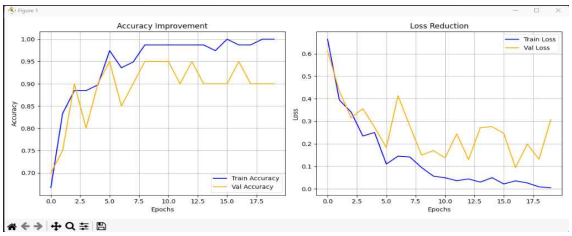

그림 2. 데이터 증강 적용 전 학습 곡선 (정확도, 손실값)
Fig. 2. Learning Curve before Data Augmentation (Accuracy, Loss)

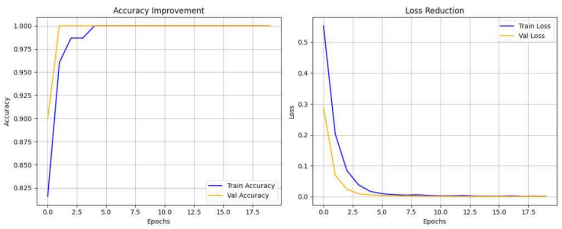

그림 3. 데이터 증강 적용 후 학습 곡선 (정확도, 손실값)
Fig. 3. Learning Curve After Data Augmentation (Accuracy, Loss)

5. 실험 결과 및 분석

5.1 시각적 발화 감지 모델 성능

데이터 증강 기법을 적용한 결과, 초기 발생하던 과적합 문제가 해소되고 안정적인 학습 곡선을 확인할 수 있었다. [그림 3]와 같이 학습이 진행됨에 따라 검증 데이터셋(Validation Set)에 대한 정확도(Accuracy)가 1.0(100%)에 수렴하였으며, 검증 손실(Loss) 또한 0.0004 수준으로 매우 낮게 기록되어 모델이 발화 의도를 명확하게 구분함을 확인하였다. 총 20개의 검증 샘플에 대하여 모델은 모든 케이스를 정확하게 분류하였다. 특히, 사용성 관점에서 가장 중요한 '발화(Speaking)' 클래스에 대해 100%의 재현율(Recall)을 달성하였다. 이는 키오스크 시스템이 사용자의 발화 시도를 누락 없이 감지할 수 있음을 시사하며, 실제 서비스 적용 시 교통 약자의 음성 명령 인식 실패율을 최소화할 수 있음을 입증한다.

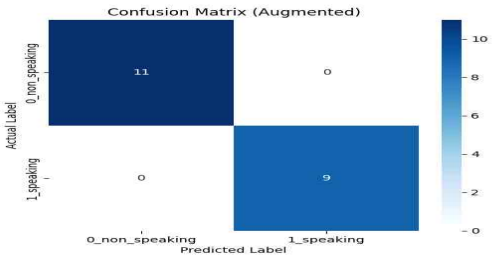

그림 4. 학습 모델의 검증 데이터셋에 대한 오차 행렬
Fig. 4. Error Matrix for Validation Data Set of Learning Model

5.2 파이프라인 버전별 비교

본 연구에서는 최적의 파이프라인을 도출하기 위해 여러 버전의 실험을 수행하였다. V1(Sepformer 무조건 적용)의 경우, 한국어 음성 신호가 심각하게 왜곡되어 베이스라인 대비 WER이 17.46%p 증가하였다. 이는 딥러닝 기반 고성능 모델의 무분별한 적용이 오히려 역효과를 초래할 수 있음을 실증적으로 보여준다.

반면, 제안하는 V6 파이프라인은 SNR 6dB 임계값을 기준으로 조건부 처리를 적용하여 전체 WER을 1.07%p 감소시키는 성능 개선을 달성하였다.

표 2. 파이프라인 버전별 성능 비교(V1~V6)
Table 2. Comparison among Pipeline Versions

버전	전처리 방식	SNR 임계값	노이즈 제거 모델	Skip률	WER 변화	결과
Base line	없음 (원본)	-	-	0%	기준 (0.0%p)	Whisper 단독
V1	무조건 적용	없음	Sepformer	0%	-17.46%p	한국어 음성 왜곡
V2	무조건 적용	없음	noisereduce (다단계)	0%	-0.25%p	거의 중립
V3	무조건 적용	없음	noisereduce + SRT 정밀 매칭	0%	-0.96%p	약간 악화
V4	조건부 적용	≥15dB Skip	noisereduce	약 50%	-0.48%p	약간 악화
V5	조건부 적용	≥8dB Skip	noisereduce	약 70%	+0.04%p	거의 중립
V6	조건부 적용	≥6dB Skip	noisereduce	79.4%	+1.07%p	성공

5.3 청각적 노이즈 제거 성능 (Audio Pipeline Evaluation)

제안하는 적응형 SNR 오디오 파이프라인(V6)의 성능을 검증하기 위해 Traffic, Construction, Complex 등 6가지 소음 환경에서 실험을 수행하였다. 실험 결과, 기존 딥러닝 모델(Sepformer)을 무조건적으로 적용했을 때는 한국어 음성 신호 왜곡으로 인해 WER(단어 오류율)이 17.46%p 증가하는 심각한 성능 저하가 발생하였다.

표 3. V6 소음 환경별 파이프라인 상세 결과
Table 3. Pipeline Results against Noise Types

환경	평균 SNR	처리 방식	Skip률	WER 변화
Traffic	4.7 dB	노이즈 제거	17%	+3.78%p (개선)
Complex	6.1 dB	혼합	67%	+2.63%p (개선)
Construction	16.8 dB	Skip	100%	0.00%p (유지)
Factory	18.8 dB	Skip	100%	0.00%p (유지)
Facility	18.3 dB	Skip	100%	0.00%p (유지)
Outdoor	8.3 dB	Skip	100%	0.00%p (유지)
평균	12.3dB	-	79.4%	+1.07%p

반면, 제안하는 SNR 6dB 임계값 기반의 적응형 파이프라인(V6)은 전체 데이터의 79.4%에 대해 처리를 생략(Skip)하여 음성 본연의 정보를 보존하면서도, 전체 WER을 1.07%p 감소(성능 개선)시키는 성과를 거두었다. 특히 노이즈 제거가 필수적인 저SNR 환경(6dB 미만)에서는 평균 3.21%p의 높은 WER 감소 효과를 보여, 시스템이 소음 환경에 유연하고 효과적으로 대응함을 입증하였다.

5.4 효율성 및 연산 비용 분석

본 연구에서 제안한 조건부 처리 방식(V6)의 실용성을 검증하기 위해 연산 효율성과 성능 유지 여부를 분석하였다. 분석 결과, 79.4%의 높은 Skip률은 다음 세 가지 측면에서 시스템의 현장 적용성을 입증한다. 첫째, 연산 속도의 향상이다. noisereduce 기반의 노이즈 제거 과정은 30초 오디오 청크당 평균 0.12초의 처리 시간이 소요된다. 제안 모델은 전체 데이터의 79.4% 구간에서 이 과정을 생략함으로써 전처리 연산량을 획기적으로 절감하였다. 둘째, 에너지 효율성이다. 불필요한 연산의 생략은 엣지 디바이스(Edge Device)의 전력 소모를 감소시켜, 24시간 운영되는 키오스크의 운영 비용 절감에 기여한다. 셋째, 성능의 안정성이다. 고SNR 환경(6{dB} 이상)에서 처리

를 생략하였음에도 불구하고, 해당 구간의 WER 변화는 0.00%로 나타났다. 이는 Whisper 모델의 자체 잡음 내성을 활용하여 불필요한 왜곡을 방지하면서도 인식 품질을 완벽히 유지했음을 보여주는 결과이다.

5.5 통계적 분석 및 결과 해석

V6 파이프라인의 성능을 통계적으로 검증하기 위해 67개 청크에 대한 paired t-test를 수행하였다. 전체 데이터에 대해 베이스라인 대비 평균 0.30±2.72%p의 WER 감소가 관찰되었으나, 통계적 유의성은 확인되지 않았다(t=0.906, p=0.368, 95% CI: [-0.36, 0.97]%p). 그러나 이 결과는 본 연구의 "Do No Harm" 원칙이 의도대로 작동하였음을 보여준다. 전체 청크의 79.4%(54개)가 SNR ≥ 6dB 조건을 충족하여 노이즈 제거를 생략(Skip)하였고, 이 구간에서는 WER 변화가 0.00%p로 원본 품질이 완벽히 보존되었다. 이는 고SNR 환경에서 불필요한 전처리로 인한 품질 저하를 원천적으로 방지하였음을 의미한다. 노이즈 제거가 실제로 적용된 저SNR 환경(유효 GT 기준 13개 청크, 평균 SNR 4.7dB)에서는 평균 1.56±6.01%p의 WER 개선이 관찰되었다. 특히 Traffic 환경(평균 SNR 4.7dB)에서 53.3%→51.9%로 1.33%p의 개선을, Complex 환경(평균 SNR 6.1dB)에서 42.8%→42.3%로 0.47%p의 개선을 달성하였다. 다만, 저SNR 청크의 표본 크기가 13개로 제한되어 개별적인 통계적 유의성은 확보되지 않았다(p=0.385).

본 연구의 핵심 기여는 전체 WER 개선량의 극대화가 아니라, 불필요한 처리로 인한 품질 저하를 방지하면서 저SNR 환경에서만 선별적으로 개입하는 실용적 파이프라인의 제안에 있다. 통계적 유의성의 부재는 79.4%의 청크가 처리를 생략하여 WER 변화가 0인 점에 기인하며, 이는 보수적 설계 철학의 자연스러운 결과이다.

6. 결론 및 향후연구

본 연구는 교통 약자의 디지털 사이니지 사용성 향상을 위한 음성-영상 융합형 UI 파이프라인을 설계하고, 그 효과를 실험을 통해 통계적으로 제시하였다.

첫째, 데이터 증강(Data Augmentation)과 전이 학습을 적용한 MobileNetV2-GRU 기반 발화 감지 모델은 100%의 정확도와 재현율을 달성하였다. 이는 데이터가 부족한 환경에서도 시각적 모듈이 사용자의 발화 의도를 누락 없이(Zero False Negative) 감지하여, 디지털 기기의 활용에 제약이 있는 사용자에게 신뢰성 높은 인터페이스를 제공할 수 있음을 입증하였다. 둘째, 순환 버퍼(Circular Buffer) 메커니즘을 통해 음성 인식의 고질적인 문제인 '초기 발화 손실'을 원천적으로 차단하였다. 시각적 발화 감지 시점 이전의 3초간 데이터를 활용함으로써, 사용자가 호출어 없이 자연스럽게 말을 시작하더라도 온전한 음성 명령 처리가 가능해졌다. 셋째, 딥러닝 기반 노이즈 제거 모델(Sepformer)은 한국어 음성 신호를 왜곡하여 오히려 성능을 저하시키는 한계가 있음을 규명하였다. 넷째, SNR 6dB 임계값을 적용한 적응형 처리 전략은 전체 데이터의 79.4%를 처리 생략하면서도 WER을 1.07%p 감소시키는 효율적인 결과를 달성하였다. 다섯째, 저SNR 환경에서만 선별적 노이즈 제거를 적용할 때 평균 3.21%p의 높은 WER 감소 효과를 얻을 수 있다. 본 연구의 핵심 발견은 '고성능 모델의 도입이 반드시 성능 향상을 보장하지 않는다'는 점이다. Whisper는 이미 자체 잡음 내성을 보유하고 있으며, 고SNR 환경에서는 전처리가 오히려 해로울 수 있다. 따라서 저SNR 구간에 대해서만 선별적으로 개입하는 전략이 가장 효과적인 것으로 확인되었다.

향후 연구에서는 더 다양한 사용자 환경과 동

적 임계값 알고리즘을 통해 시스템의 범용성을 확장할 필요가 있다. 또한, 실제 사용환경에서는 잡음의 환경이 다양하여 동적 잡음환경에 대응하기 위해 실시간으로 SNR 임계값을 조정하는 적응형 알고리즘으로 변화가 필요하다. 발화구간 추정에서도 입술 외에도 눈썹의 움직임, 머리의 미세한 진동 등 보조적인 시각적 특징을 융합하는 멀티모달(Multimodal) 분석을 도입하여 입이 가려진 환경에서도 청각적 VAD 모드로 즉시 전환(Fallback)하는 하이브리드 로직을 추가하여 강인성을 확보할 수 있다.

본 연구결과를 실제 키오스크에 적용하기 위해서는 몇 가지 실용적 고려사항이 존재한다. 먼저, 하드웨어 측면에서 Whisper large-v3 모델은 약 4.5GB의 GPU 메모리를 요구하므로, 엣지 디바이스 배포 시 경량화된 모델(small, medium)로의 대체 또는 클라우드 기반 처리 방식을 고려해야 한다. 또한, 개인정보 보호 관점에서 얼굴 영상과 음성 데이터의 처리 및 저장에 관한 법적 요건을 충족해야 하며, 가능한 한 온디바이스(on-device) 처리를 통해 데이터 유출 위험을 최소화하는 것이 바람직하다.

표 4. 최종 시스템 성능 요약
Table 4. Performances for Final Pipeline System

평가 지표	결과	달성 지표값 설명
발화 감지 정확도	100%	20/20 검증 샘플 정확 분류
발화재현율	100%	음성 명령 누락 없음
오디오 처리 생략률	79.4%	고SSNR 데이터 보존
전체 WER 개선	+1.07%p	Baseline 모델 대비
저 SNR WER 개선	+3.21%p	Traffic + Complex 평균

<Acknowledgement>
“본 연구는 2023년도 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업 지원을 받아 수행되었음(2023-0-00089127821267501O3).”

참고문헌

[1] National Information Society Agency, “2024 The Report on the Digital Divide”, 2024, DOI: https://www.nia.or.kr/site/nia_kor/ex/bbs/View.do?cbIdx=81623&bcIdx=27832

[2] Korea Consumer Agency, “Survey on the use of kiosks”, 2023. DOI: https://www.kca.go.kr/smartconsumer/sub.do?menukey=7301&mode=view&no=1003409523

[3] Msweli, N. T., & Mawela, T.“ Financial inclusion of the elderly: Exploring the role of mobile banking adoption”. Acta Commercii, 20(1), pp. 1-10,. 2020, DOI: 10.18267/j.aip.143

[4] Brunner, M., Hemsley, B., Togher, L., & Palmer, S., “Technology and its role in rehabilitation for people with cognitive-communication disability following a traumatic brain injury”. Brain Injury, 31(8), pp. 1028-1043, 2017, doi: 10.1080/02699052.2017.1292429

[5] National Information Society Agency, “Kiosk UI platform Guideline”, 2021. DOI: https://www.nia.or.kr

[6] Tao, R., Pan, Z., Das, R. K., Qian, X., Shou, M. Z., & Li, H., “Is someone speaking? Exploring long-term temporal features for audio-visual active speaker detection”. Proc. of the 29th ACM International Conference on Multimedia. pp. 3927-3935, 2021., https://doi.org/10.48550/arXiv.2107.06592

[7] Chung, J. S., & Zisserman, A., “Lip reading sentences in the wild”. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6447-6456, 2017, https://doi.org/10.48550/arXiv.1611.05358

[8] Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C. L., Yong, M. G., Lee,

J., & Grundmann, M., “MediaPipe: A framework for building perception pipelines”. arXiv preprint arXiv:1906.08172. 2019.
[9] Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., & Zhong, J. “Attention is all you need in speech separation”. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 21-25, 2021, https://doi.org/10.48550/arXiv.2010.13154
[10] Luo, Y., & Mesgarani, N., “Conv-TasNet: Surpassing ideal time - frequency magnitude masking for speech separation”. ACM Transactions on Audio, Speech, and Language Processing, 27(8), pp. 1256-1266, 2019., https://doi.org/10.48550/arXiv.1809.07454
[11] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I., “Robust speech recognition via large-scale weak supervision”. International Conference on Machine Learning, pp. 28492-28518, 2023, https://doi.org/10.48550/arXiv.2212.04356
[12] Défossez, A., Usunier, N., Bottou, L., & Bach, F. “Music source separation in the waveform domain”. arXiv preprint arXiv:1911.13254., 2019
[13] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M., “wav2vec 2.0: A framework for self-supervised learning of speech representations”. Advances in Neural Information Processing Systems, 33, pp.12449-12460, 2020, arXiv:2006.11477v3
[14] Gulati, A., Qin, J., Chiu, C. C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., & Pang, R., “Conformer: Convolution-augmented transformer for speech recognition”. Interspeech 2020, pp. 5036-5040, 2020, arXiv:2005.08100v1
[15] Loizou, P. C., “Speech Enhancement: Theory and Practice”, CRC Press, 2013, ISBN:978-1-4665-0421-9
[16] Bengio, Y., Léonard, N., & Courville, A., “Estimating or propagating gradients through stochastic neurons for conditional computation”. arXiv preprint arXiv:1308.3432.2013
[17] Yosinski, J., Clune, J., Bengio, Y., & Lipson, H., “How transferable are features in deep neural networks?”, Advances in Neural Information Processing Systems, 27, pp. 3320-3328. 2014, arXiv:1411.1792v1

저 자 소 개

김정현(Jeong-Hyun Kim)

2021년~현재 신한대학교 소프트웨어학과 재학
<주관심분야> 인공지능, 음성 잡음 제거

박종섭(Jong-Seob Park)

2021년~현재 신한대학교 소프트웨어학과 재학
<주관심분야> 인공지능, 음성 잡음 제거

저 자 소 개

최용수(YongSoo CHOI)

1998년 강원대학교 제어계측공학과 공학사
2000년 강원대학교 제어계측공학과 공학석사
2006년 강원대학교 제어계측공학과 공학박사
2006년～2007년 연세대학교 첨단융합건설 연구단 연구교수.
2007년～2013년 고려대학교 정보보호대학원 연구교수.
2013년～2020년 성결대학교 파이데이아대학 (멀티미디어) 조교수
2020년~ 현재 신한대학교 미래자동차공학과 부교수
<주관심분야> Digital Forensics, Information Hiding, Multimedia Watermarking, Steganography