논문 2025-4-20 http://dx.doi.org/10.29056/jsav.2025.12.20

저사양 엣지 환경을 위한 시각 정보 기반의 효율적 잡음 제거 및 발화 감지 파이프라인

고현지*, 서원후*, 최용수**†

Enhancing Voice Recognition Accuracy through Sequential Application of Visual Speech Detection and Noise Reduction

Hyeon-Ji Ko*, Won-Hu Seo*, Yong-Soo Choi**†

요 약

기존의 에너지 기반 VAD(Voice Activity Detection)는 소음이 심한 환경에서 음성과 비음성을 명확하게 구분하지 못하고, 불필요한 연산을 유발하는 한계를 가진다. 본 논문은 저사양 엣지 디바이스 환경에서 배경 소음으로 인해 음성인식률이 저하되는 문제를 해결하기 위한 전처리 파이프라인으로 시각적 발화 검출(Visual Voice Activity Detection)을 시스템의 연산 차단 게이트(Gate)로 사용하는 4단계 순차적 파이프라인을 제시한다. 제안하는 시스템은 파이프라인에서 검증된 발화 구간에 대해서만 선별적으로 잡음 제거 및 STT를 수행함으로써, 딥러닝 기반의 모델 없이도 기초적인 알고리즘 조합만으로 실시간성을 확보하고 비발화 구간의 잡음을 효과적으로 차단할 수 있음을 보였다. 실험 결과, RTF 0.134를 유지하면서 발화 구간에서의 잡음제거성능을 15.67dB로 향상시켰으며, 배경 잡음이 제거됨에 따라 Speech Loss는 14.59dB로 나타나 일반적인 잡음제거 VAD보다 전반적으로 개선된 성능을 보였다.

Abstract

Traditional energy-based voice activity detection (VAD) fails to clearly distinguish between speech and non-speech in noisy environments and leads to unnecessary computations. It presents a pre-processing pipeline to address the issue of decreased speech recognition accuracy due to background noise in low-spec edge device environments. This study designs a four-stage sequential pipeline that uses visual voice activity detection as a computational gating mechanism in the system. The proposed system selectively performs noise reduction and STT only on verified speech segments, demonstrating that real-time performance can be achieved and non-speech noise can be effectively blocked using only a basic combination of algorithms, without the need for deep learning-based models. Experimental results show that the system maintains an RTF of 0.134 while improving noise reduction performance in speech-active segments to 15.67 dB, and as background noise is removed, a Speech Loss of 14.59 dB is observed, demonstrating overall improved performance compared to conventional noise-removal-based VAD.

한글키워드 : 영상 VAD, 노이즈 차감, STT 정확도, 저사양 고성능, 디지털 사이니지

keywords : Video VAD, noise reduction, STT accuracy, high performance on low-spec, digital signage

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

접수일자: 2025.12.05. 심사완료: 2025.12.15.
게재확정: 2025.12.20.

1. 서 론

최근 비대면 서비스의 확산과 함께 공공장소의 디지털 사이니지와 키오스크는 정보 전달 매체 역할에 더불어 사용자 편의성을 높이기 위한 지능형 인터페이스로 진화하고 있다. 특히 고령자, 문해력이 낮은 이용자 등 이른바 디지털 소외 계층에게는 복잡한 메뉴 탐색이나 정교한 터치 조작이 큰 장벽이 될 수 있어, 공공 서비스 접근성을 보장하기 위한 보완 인터페이스의 필요성이 더욱 커지고 있다. 대표적인 사례로 터치 인터페이스의 접근성 한계를 보완하기 위한 음성 상호작용 기능의 도입이 있다. 그러나 버스 정류장, 음식점, 쇼핑몰 등 디지털 사이니지가 실제 운용되는 환경은 예측 불가한 배경 소음과 다수의 발화자가 섞이는 경우가 빈번히 발생한다. 이러한 요인은 STT(Speech-to-Text) 엔진의 인식률을 저하시키는 주된 원인이 된다. 음성신호의 에너지를 분석하여 발화 구간만을 검출하는 VAD(Voice Activity Detection)가 전처리 단계에서 널리 사용되지만, 이는 신호 대 잡음비(SNR)가 낮은 환경에서 치명적인 한계를 보인다[1]. 특히 주변의 웅성거리는 말소리 등의 비정상(Non-stationary) 소음을 사용자의 발화로 오인하는 오류가 잦다.

VAD의 이러한 오작동은 후속 STT 엔진에 부정확한 데이터를 전달하여 불필요한 연산 부하를 가중시킨다. 대부분의 디지털 사이니지 하드웨어는 고성능 GPU 서버 대신 저전력 임베디드 시스템 또는 보급형 PC 기반으로 구동되기 때문에, 이러한 연산 부하는 전체 시스템의 응답 속도를 늦추고 사용자 경험(UX)을 저해하는 요소로 작용한다. 시스템 응답 지연이나 반복 인식 실패는 디지털 소외 계층에게 서비스 이용 포기로 직결될 가능성이 높아, 기술적 성능 문제를 넘어 공공 서비스의 형평성과 접근성 측면에서 중요한 문제로 이어진다. 따라서 소음이 심한 환경에서도 실제 사용자의 발화만을 정확히 선별하여, 한정된 하드웨어 자원을 효율적으로 배분할 수 있는 기술이 요구된다.

최근 연구들에 따르면, 입술 움직임과 같은 시각 정보를 음성 처리와 같이 사용할 경우 저자원 및 고소음 환경에서도 인식 성능이 개선됨이 입증된 바 있다[2, 3].본 논문은 소음에 취약한 오디오 의존적 VAD를 배제하고, 소음에 영향을 받지 않는 시각적 발화 검출(Visual Voice Activity Detection)을 시스템의 주 게이트(Gate)로 사용하는 새로운 멀티모달 전처리 파이프라인을 제안한다. 제안하는 시스템은 카메라를 통해 사용자의 입술 움직임을 포착하여 '실제 발화(Actual Speech)' 구간을 1차적으로 판별한다. 이후 해당 구간의 오디오 데이터만을 선별하고, 비발화 구간에서 학습된 노이즈 프로파일을 기반으로 순차적인 잡음 제거를 수행하여 STT 엔진에 전달하는 구조를 가진다.

추가적인 파이프라인의 적용을 통해 두가지 이점을 얻고자 한다. 첫째, 시각 정보를 음성인식 모듈 연산의 트리거로 활용함으로써 주변 소음이 아무리 심하더라도 발화 시작점을 놓치거나 오인식하지 않는 강인성을 확보한다. 둘째, 비발화 구간에서는 무거운 오디오 처리 및 STT 연산을 원천적으로 차단(Gating)함으로써, 한정된 하드웨어 자원 내에서도 실시간성을 보장하는 높은 연산 효율성을 달성한다. 결과적으로 본 연구는 공공장소의 디지털 사이니지 환경에서 음성 기반 인터페이스의 반응성을 높여, 디지털 소외 계층을 포함한 다양한 이용자가 보다 동등하게 공공 서비스를 이용할 수 있어 접근성을 향상시킬 것으로 기대한다.

2. 관련 연구

2.1 오디오 기반 잡음 제거 연구

전통적인 음성 향상을 위한 기술은 스펙트럼 차감법이나 위너 필터와 같은 통계적 모델에 기반하였다[4]. 이러한 방식은 연산량이 적어 실시간 처리에 적합하지만, 비정상 잡음이 있는 환경에서는 음성 신호까지 훼손해버리는 뮤지컬 노이즈(Musical Noise) 현상이 발생하기 쉽다는 한계가 지적되어 왔다[5]. 최근 딥러닝의 발전으로 심층 신경망(DNN)이나 순환 신경망(RNN), Transformer 구조를 활용한 잡음 제거 기술이 주를 이루고 있다. 그러나 딥러닝 모델들은 수백만 개 이상의 파라미터를 가지며, 실시간 추론을 위해서는 고성능 GPU 연산 자원을 필요로 한다. 이는 키오스크나 디지털 사이니지와 같은 저전력 임베디드(Edge Device) 환경에서 높은 비용과 발열, 지연 시간(Latency)을 동반하는 문제점이 있다[6]. 이러한 한계로 인해, 오디오 처리 복잡도를 높이지 않으면서도 잡음 환경에서의 인식 성능을 보완할 수 있는 보조 정보에 대한 연구가 필요하게 되었다.

2.2 시각 정보를 활용한 연구

앞서 언급한 오디오 기반 처리의 한계를 극복하기 위해 입술 움직임(Lip reading)과 같은 시각 정보를 결합하는 시청각 음성 인식(Audio-Visual Speech Recognition, AVSR) 연구가 대안으로 주목받고 있다. 시각 정보는 주변 잡음의 영향을 전혀 받지 않으므로 SNR이 낮은 고소음 환경에서 오디오 처리에 도움이 된다는 것이 증명되었다[7, 8]. Google의 'Looking to Listen'[9] 연구는 시각과 청각의 정보를 사용하며 화자를 분리하는 성능을 높였으나, 영상 처리와 음성 처리를 위한 두 개의 신경망을 동시에 실행해야 하므로 연산 복잡도가 매우 높다. 최근에는 경량화를 위한 연구가 진행되고 있으나[10], 제한된 하드웨어에서 실시간으로 구동하기에는 여전히 큰 부담이다. 따라서 본 연구는 무거운 모델을 사용하는 대신, 시각 정보를 게이팅(Gating) 트리거로만 활용하여 연산 효율성을 높이고 엣지 환경에서의 실용성을 확보한다.

3. 제안하는 발화구간 탐지 및 잡음제거 시스템

제안시스템(VVAD: Visual Voice Activity Detection)은 그림1에서와 같이 Video모듈이 1차적으로 발화여부를 판별하면, 발화 구간의 음성 신호가 Audio모듈의 잡음 제거 로직을 거쳐 최종적으로 STT 엔진에 정제된 오디오를 전달하는 4단계 순차적 구조를 가진다.

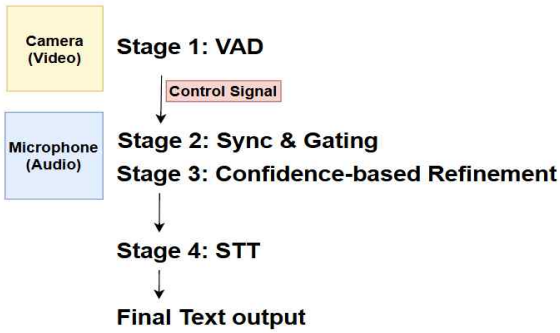

그림 1. 4단계 시각적 VAD 흐름도
Fig. 1. 4-Step Visual VAD flowchart

3.1 1단계: 시각적 발화 검출

1단계는 카메라 입력을 활용해 오디오를 처리하기 전에 사용자가 실제로 말을 하고 있는지 판단하는 '시각적 게이트(Visual Gate)' 역할을 한다. 기존의 청각 기반 VAD는 소음이 큰 환경에서는 잡음을 발화로 잘못 인식하는 문제가 있다. 이를 해결하기 위해 본 시스템은 소음의 영향을 받지

않는 시각 기반 발화 검출(Visual VAD) 기법을 도입했다. Visual VAD 과정은 실시간 사용자 탐지, 입술 특징 추출, 그리고 시각적 신뢰도(Visual Confidence) 계산의 세 단계로 이루어진다.

공공장소는 여러 사람이 오가는 환경이기 때문에, 시스템과 실제로 상호작용하려는 사용자를 구분하는 과정이 필요하다. 본 시스템은 사용자 식별을 위해 두 가지 환경을 정의한다. 첫째, 사용자는 기기 정면에 위치해 화면을 바라보고 있다. 둘째, 주변에 다수의 사람이 있더라도 실제로 기기와 상호작용하는 발화자는 한 명이다. 사용자 식별에는 YOLOv5n 모델을 사용했다[11]. 영상 입력이 들어오면 YOLOv5n은 프레임 내의 모든 사람 객체를 검출하고, 여러 사람이 감지될 경우, 앞선 설정한 전제에 따라 카메라에 가장 가까워 상호작용 가능성이 높은 사람을 주 발화자로 판단하다. 이를 위해 바운딩 박스(Bounding Box) 면적이 가장 큰 사람을 선택해 얼굴 영역의 ROI(Region of Interest)를 최종적으로 확정한다.

선정된 ROI 내에서 입술의 기하학적 특징을 정량적으로 추출하기 위해 Google의 MediaPIpe Face Mesh를 적용하였다. MediaPIpe Face Mesh는 단일 카메라 입력만으로도 얼굴 표면의 3D 구조를 추정할 수 있으며, 총 468개의 고유한 얼굴 랜드마크(Landmark) 좌표 x, y, z 값을 반환한다[12]. 단순 2D 좌표만을 사용할 경우 사용자가 측면을 바라보거나 고개를 회전할 때 입술 형태가 영상 평면에서 왜곡되어 오탐지 발생 가능성이 존재한다. 이를 보완하기 위해 본 연구에서는 MediaPIpe가 제공하는 3D 랜드마크 정보를 활용하여 얼굴의 회전 각도를 추정하였다. 특히 횡축(Yaw) 회전값이 설정된 임계값(Threshold)을 초과할 경우, 사용자가 디스플레이를 응시하지 않는 것으로 판단하여 비발화 상태(non-speech state)로 처리하였다. 이러한 절차는 사용자가 사이니지를 바라보지 않는 상황에서 발생할 수 있는 불필요한 발화 판단을 효과적으로 억제한다.

본 연구에서는 발화 여부를 판단하는 알고리즘으로 CNN(Convolutional Neural Network)이나 LSTM과 같은 고연산 딥러닝 모델 대신, 입술의 기하학적 특성에 기반한 규칙 기반(Rule-based) 접근법을 채택하였다. 최근 시각적 음성 인식 분야에서는 고성능 딥러닝 모델을 활용하는 사례가 증가하고 있으나, 이러한 모델은 엣지 컴퓨팅 환경에서 상당한 연산 자원과 메모리 대역폭을 요구한다[6]. 본 시스템은 디지털 사이니지 환경에서 STT 엔진과 병렬적으로 구동되어야 하므로, 사용자 경험(UX)을 저해하지 않는 실시간 반응성(Real-time Responsiveness) 확보가 핵심 요구사항이다[13]. 이에 따라 본 연구는 연산 부담이 적고 처리 속도가 빠른 규칙 기반 방식을 적용하여 시스템 지연 시간(Latency)을 최소화하였다. 이러한 경량화된 설계에 따라, 발화 판단을 위한 특징 벡터로는 입술 종횡비(LAR, Lip Aspect Ratio)와 프레임 간 차분(Frame Difference)을 결합하여 사용하였다. LAR은 단독으로는 입을 벌린 채 정지한 상태를 정확히 구분하기 어렵고, 프레임 차분은 조명 변화나 미세한 움직임에 취약하다는 한계가 있다. 따라서 두 특징을 상호보완적으로 결합함으로써 각각의 단점을 효과적으로 상쇄하였다.

첫째, 입술 종횡비(LAR)는 입술의 기하학적 개폐 정도를 수치적으로 표현하는 지표로, 그 정의는 (1)과 같다. 여기서 사용되는 좌표 P_i는 MediaPipe가 제공하는 정규화 좌표계(Nomalized Coordinate System)를 기반으로 하며, i는 특정 랜드마크의 인덱스를 나타낸다. LAR 산출에는 입술 외곽을 구성하는 주요 랜드마크 집합을 사용하여 발화 관련 형태적 변화를 정밀하게 포착하였다. 구체적으로, 윗입술 상단 랜드마크(Index: 13)와 아랫입술 하단 랜드마크(Index:

14), 그리고 좌우 입꼬리 랜드마크(Index: 61, 291)의 좌표를 이용하여 각 지점 간 유클리드 거리(Euclidean Distance)를 산출한다. 이러한 정규화 좌표 기반 비율 계산 방식은 얼굴 크기나 카메라와의 거리 변화에 영향을 받지 않기 때문에, 사용자와의 상대적 거리 차이가 커도 일관된 성능을 보장한다. LAR 값은 입이 벌어질수록 증가하며, 계산된 값이 설정된 임계값을 초과할 경우 개구(Open) 상태로 판단한다.

$$LAR = \frac{|p_{top} - p_{bottom}|}{|p_{left} - p_{right}|} \quad (1)$$

둘째, 발화 시 동반되는 동적인 움직임을 포착하기 위해 이전 프레임과 현재 프레임의 입술 ROI 영역에 대해 픽셀 단위 변화량을 계산하여 프레임 간 차분(Frame Difference) 특징을 추출하였다. 본 연구에서는 오탐지와 미탐지 사이의 균형을 유지하기 위해, 사전 실험을 통해 표 1과 같은 임계값을 설정하였다. 임계값 설정에서 가장 중점을 둔 부분은 발화 구간의 보존이다. 임계값을 지나치게 엄격하게 설정하면, 실제로 입을 입을 작게 벌리고 말하는 구간이 비발화로 잘못 분류되는 미탐지(False Negative)가 발생할 수 있다. 1단계에서 비발화로 차단된 신호는 이후 단계에서 복구가 불가능하므로, 본 연구는 실제 발화를 놓치지 않도록 비교적 여우 있는 임계값을 채택하였다.

표 1. 시각적 발화 검출을 위한 임계값 설정
Table 1. Threshold Settings for Visual Speech Detection

Parameter	Value
LAR 임계값	0.30
움직임 임계값	0.02
자세 허용 각도	0.35
연속 프레임 수	2

단일 프레임에 기반한 판단은 일시적인 입술 움직임이나 영상 노이즈에 취약하기 때문에, 본 연구에서는 검출된 특징들을 종합하여 발화 신뢰도(Confidence Score)를 산출하였고 계산식으로 표현하면 (2)와 같다. 여기서 입술이 열린 비율에 따른 신뢰도 점수(S_{ratio})는 조건식에 의해 결정된다. 신뢰도 산출은 움직 점수(S_{diff})과 입술 형태 점수(S_{ratio})에 각각 가중치(W_1 = 0.4, W_2 = 0.6)를 부여하여 계산되며, 이는 고개 끄덕임과 같은 비발화성 움직임을 발화로 오인하는 문제를 완화한다. 또한 발화 과정에 나타나는 짧은 묵음을 자연스럽게 보정하기 위해 이전 프레임의 신뢰도를 70% 반영하는 지수 이동 평균(EMA, Exponential Moving Average) 기반의 스무딩(α = 0.3)을 적용하였다. 이 과정을 통해 (3)의 최종 발화 신뢰도(C_{final})가 임계값 이상일 때 발화(Speaking) 상태로 판단한다. 최종적으로 구성된 결과는 표 2의 딕셔너리 형태로 오디오 모듈에 전달되며, 이 중 timestamp 필드는 오디오 버퍼와의 정확한 시간 축 정렬(Time-Alignment)을 보장하는 핵심적인 역할을 수행한다.

$$S_{diff} = \min(Diff \times 5.0, 1.0)$$
$$S_{ratio} = \begin{cases} \min(5.0(LAR - 0.2), 1.0) & \text{if LAR} > 0.2 \\ 0 & \text{otherwise} \end{cases}$$
$$C_{raw} = w_1 \cdot S_{diff} + w_2 \cdot S_{ratio} \quad (2)$$

$$C_{final_t} = \alpha \cdot C_{raw} + (1 - \alpha) \cdot C_{final_{t-1}} \quad (3)$$

3.2 2단계: 발화영역 오디오 청크 분리 및 기본 잡음 제거

2단계는 모든 유효 청크에 대해 수행되는 1차 억제 단계로, 비발화 구간을 이용한 노이즈 프로파일링 추정·갱신과 이를 기반으로 하는 스펙트럼 감산을 통해 SNR을 끌어올리는 데에 목적을 둔다.

표 2. 시각적 발화 검출의 출력 데이터 명세
Table 2. Output Data Specifications for Visual Speech Detection

Key	Data Type	Description
frame_id	Integer	영상 스트림의 순차적 고유 프레임 번호
timestamp	Float	시스템 시작 시점 기준 경과 시간 (단위: sec)
roi	Dictionary	검출된 입술 영역의 좌표 정보 {x, y, w, h}
is_speaking	Boolean	시각 정보 기반의 최종 발화 판정 결과 (True/False)
confidence	Float	발화 가능성에 대한 확률적 신뢰도 점수 (0.0 ~ 1.0)
person_ detected	Boolean	사용자 객체 탐지 성공 여부

엔트로피 기반 및 에너지 기반 결합 지표는 고잡음 환경에서 발화 탐지의 보완책으로 제시되어 왔다[14]. 이에 따라 2단계는 1단계의 시각적 판별 정보(타임 스탬프)를 받아 오디오 스트림과 동기화하고, 기본적인 잡음 제거를 수행한다. 시스템은 1단계의 시각정보와 오디오 버퍼의 타임 스탬프 간의 차이(Δt)를 계산하여 동기화를 수행하며, 허용 오차범위(Sync Tolerance) 이내의 오디오 청크만을 유효 처리 대상으로 한다.

speech active state(is_speaking)가 True로 판별된 청크(발화 구간)는 노이즈 감산 로직을 통해 1차적으로 정제된다. 반면, False로 판별된 청크(비발화 구간)은 시스템의 연산 부하를 줄인 후 복원 과정의 소음 생성을 방지하며, 노이즈 프로파일을 학습하는 자료로 활용한다. 시스템은 비발화 구간의 오디오 스펙트럼을 분석하여 배경 소음의 특성을 실시간으로 업데이트한다. 수집된 비발화 프레임들의 평균 스펙트럼(μ_{noise})은 (4)와 같이 학습률 α(update factor)에 따라 기존 프로파일($N_{profile}$)에 반영된다.

$$N_{profile}^{(t)} = (1-\alpha) \cdot N_{profile}^{(t-1)} + \alpha \cdot \mu_{noise} \quad (4)$$

이 방식은 stationary=False 설정과 결합되고, 시스템이 급격한 환경 소음 변화(in-flight noise)에도 유연히 적응할 수 있도록 한다. 특히 초기 작동 시 별도 비발화 샘플을 입력받지 못하는 환경을 고려하여 같은 설정 하에 발화 청크 내부에서 상대적으로 에너지가 낮은 성분을 소음으로 추정하는 자동 인플라이트 프로파일링을 채택한다.

또한 2단계는 프로파일 상태 머신(FSM)을 도입하여 초기화 → 안정화 → 재학습의 단계를 이행한다. 초기화 단계에서는 시스템 기동 시 짧은 기간(예: 2-5초)의 빠른 프로파일 수렴을 허용하여 기존 소음 플로어를 확보하고, 프로파일을 보정해 과도한 환경 추적을 억제한다. 만약 소음 평균치의 변화율 등의 실시간 통계가 사전 설정한 임계치를 초과하면 재학습단계로 전환하여 새 환경에 맞는 프로파일을 학습하도록 한다. 더불어 초기 콜드스타트(cold start) 안정성을 보장하기 위해 FSM은 시스템 초기화 상태를 유지하고, 최소 4프레임 이상의 비발화 데이터를 수집하여 초기 노이즈의 기준점을 잡는다. 이는 장시간 운용 시의 적응성 사이의 균형을 보장하는 핵심 매커니즘이다.

실제 연산에서는 prop_decrease(스펙트럼 감산 기반 노이즈 억제 알고리즘에서 사용되는 노이즈 감산 강도 파라미터)와 같은 강도 파라미터를 상황에 맞게 동적으로 조정한다. 0≤prop_decrease ≤1의 값을 가지며, 추정된 소음 스펙트럼에 대

해 감산 비율을 조절하는 역할을 한다. 런타임 환경은 RMS, 간이 SNR(발화 RMS-배경노이즈 RMS), 발화 비율(유효 발화 프랙션) 등의 지표로 분석되며, 일정 간격(예: 2-5초)으로 집계된 통계에 따라 prop_decrease, 스무딩 적용 여부, 재학습 트리거 민감도 등을 자동 튜닝한다. 예를 들어 평균 SNR이 낮아 시끄러운 환경일 경우 prop_decrease를 증가시켜 더 강력한 억제를 적용하고, 반대의 경우 prop_decrease를 낮춰 음질 보존을 우선시한다. 이러한 자동 튜닝은 연산 자원과 음질 사이의 실시간 트레이드오프를 시스템 차원에서 최적화한다.

적응형 노이즈 제거의 과정은 그림 2와 같으며, 해당 단계에서의 출력(stage2_output)은 이후 3단계로 전달되어 후속의 confidence 기반 정밀 후처리 여부를 결정한다.

3.3 3단계: Confidence 기반의 정밀 잡음 제거

3단계에서는 2단계의 1차 억제 결과(stage2_output) 에 대해, 전달된 Confidence 값을 바탕으로 판정된 발화 구간에 한해 선택적으로 정밀한 후처리를 수행한다. 구체적으로, Confidence가 설정한 임계값(예: 0.8) 이상인 구간에 대해 스펙트럼 도메인에서 국소저거 스무딩을 적용하며, 임계값 미만의 구간은 추가 처리 없이 stage2_output을 그대로 전달하여 불필요한 연산을 방지한다. 이러한 적응형(Adaptive)방식은 연산 비용과 오디오 품질 간의 균형을 맞추는 데에 중심을 둔다.

정말 후처리의 핵심은 두 가지 목적을 동시에 달성하는 것으로, 2차 단계에서 강력한 억제를 위해 높은 prop_decrease를 적용한 후 발생한 음성 성분 손실과 'musical noise' 계열의 인공 잡음을 완화하는 것과, 확실한 발화 구간에 대해 음성 복원을 최대화하는 것이 그 목적이다. 구체적인 처리 흐름은 다음을 따른다.

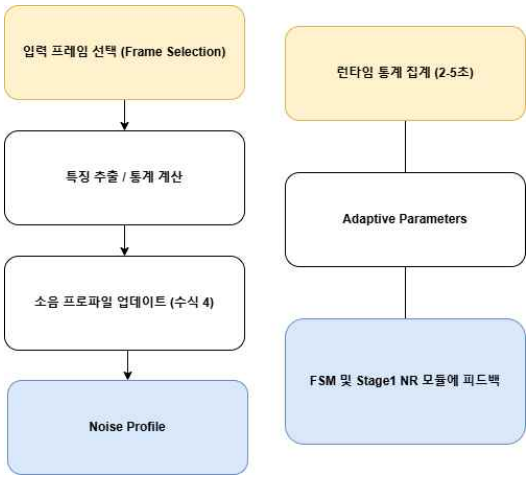

그림 2. 적응형 노이즈 프로파일링 및 자동 튜닝 프로세스

Fig. 2. Adaptive Noise Profiling and Auto-Tuning Process

(1) stage2_output을 STFT로 변환하여 스펙트로그램을 획득한다.

(2) confidence가 임계값 이상인 경우에 한해 스펙트럼 스무딩(Spectral Smoothing)을 적용한다. 본 스무딩은 2차원 미디안 필터(median filter) 또는 이와 동등한 로컬 보간 기법을 사용하여 스펙트럼 감산 과정에서 발생할 수 있는 시간-주파수 영역의 국소적 불연속(local spectral discontinuity)과 산발적 잔류 성분을 완화한다. 이 과정은 musical noise 제거와 동시에 인위적 스펙트럼 왜곡을 줄인다.

(3) 필요 시, 자동 튜닝 모듈의 산출값을 반영하여 stage1 단계에서 가혹하게 낮춘 주파수 밴드의 gain을 제한적으로 보상한다(예: 특정 주파수 대역의 attenuation 한도 설정). 이 보정은 confidence 가중치에 따라 강도를 조절한다(confidence가 높을수

록 보정 폭을 넓히고, 낮을수록 보수적으로 적용한다).

(4) confidence가 임계값 미만이면 추가 처리 대신 stage2_output을 그대로 다음 단계로 전달한다(후속 단계에서의 불필요한 연산·왜곡 방지)

3단계에서는 1단계에서 산출된 발화 신뢰도(Confidence, C)를 기반으로 확실한 발화에 대한 후처리를 수행하는데, 음질의 자연스러움을 극대화하는 적응형 스무딩 (Adaptive Smoothing)이 포함된다. 수행 과정에서 모듈은 인공 잡음(Musical Noise) 억제를 위해 시스템은 신뢰도에 반비례하는 강도의 가우시안 필터를 적용한다. 가우시안 필터의 표준편차 σ는 (5)와 같이 계산한다.

$$\sigma = k \cdot (1.0 - C) \quad (5)$$

더 나아가 3단계는 2단계에서 수행되는 자동 튜닝 루프와 피드백을 주고받는다. 예컨대 3단계의 복원 성공도를 간이 메트릭(예: 스펙트럼 연속성 지수, 음성 지향의 SNR 개선량)으로 평가하여 그 결과를 2단계의 파라미터 조정(예: prop_decrease 보정, 안정화 계수 등)에 반영할 수 있다. 이로써 시스템은 단순한 상하향식 하드코드 정책이 아닌 실시간 성능 기반의 폐쇄 루프(closed-loop adaption) 적응 정책을 수행하게 된다.

3.4 실시간 처리를 위한 스트리밍 인터페이스 및 상태 관리

본 시스템은 디지털 사이니지 환경의 실시간 입력(Live Stream)처리가 요구된다. 사용자의 발화에 즉각적으로 반응해야하기 때문이다. 전체 오디오 데이터를 한 번에 메모리에 로드하여 처리하는 일괄 처리(Batch Processing) 방식은 높은 인식 정확도를 보장할지언정 발화 종료까지 기다려야 하는 구조적 한계로 응답 지연이 필히 발생한다. 또한 긴 오디오 데이터를 버퍼링하는 메모리 오버헤드는 저사양 임베디드 환경에서 시스템 불안정을 초래하기 쉽다. 이를 해결하기 위해 본 모듈에서는 청크 기반의 상태 보존형(Stateful) 스트리밍 인터페이스를 설계하여 입력과 처리가 동시에 이루어지는 파이프라인을 구축하였다. 파일 기반 처리와 스트리밍 처리의 차이점은 표 3과 같다.

표 3. 파일 기반 처리와 스트리밍 처리의 차이점
Table 3. Difference between file-based processing and streaming processing

구분	파일 기반 처리	스트리밍 처리
입력 방식	전체 오디오 로드 후 처리	작은 청크 단위 연속 입력
지연 시간 (Latency)	높음 (발화 종료 후 처리 시작)	매우 낮음 (입력 즉시 처리)
상태 관리	불필요	필수
노이즈 프로파일	전체 구간 평균으로 고정	비발화 구간마다 실시간 갱신

(1) chunk 단위 처리와 제너레이션 패턴: 시스템은 들어오는 오디오 스트림을 고정된 크기의 작은 청크로 분할 처리한다. 이때 단순히 입력을 받아 출력을 내보내는 함수형 구조는 적합하지 않다. 따라서 본 연구는 Python의 제너레이터 패턴을 적용하여 호출이 종료되어도 청크의 마지막 상태가 내부 메모리에 유지되도록 구현하였다. 이전 프레임 노이즈 프로파일 및 VAD 상태가 다음 프레임으로 연속성 있게 전달될 것을 보장하기 위함이다.

(2) 경계면 불연속성 해결(Overlap-Add): 실시간 스트리밍 처리의 가장 큰 문제점은 각 청크 사이에서 발생하는 위상 불연속성(Phase Discontinuity)과 그로 인한 틱 잡음이다. 파일 처리 방식에서는 해결이 비교적 쉬우나, 스트리밍은 미래의 데이터를 알 수 없으므로 해결이 까다롭다. 이를 해결하기 위해 본 모듈은 Overlap-Add(OLA) 기법을 변형 적용한다. 현재 처리 중인 청크의 Tail 일부를 버퍼에 저장하고, 다음 청크의 Head와 겹쳐 합산함으로써 프레임 간 연결을 매끄럽게 한다.

(3) 비동기 큐를 채택한 배압 조절 영상 처리: 조절 영상 처리와 오디오 처리는 서로 다른 주기로 데이터를 생성한다(30fps/16kHz). 본 시스템은 두 모듈 사이 스레드-안전한 Deque를 배치함으로서 버퍼링을 수행한다. 이 구조는 일시적 연산 지연이 발생해도 데이터 유실 없이 순서를 보장하며, 시스템의 부하가 임계치를 넘을 경우 오래된 프레임을 드롭하거나 바이패스하여 전체 파이프라인의 실시간성을 지키게끔 한다.

3.5 4단계: STT(Sppeech-to-Text) 적용

제안하는 시스템의 마지막 단계는, 1단계에서 수행된 시각 기반 발화 검출과 2·3단계의 다단계 잡음 제거 과정을 통해 정제된 오디오 신호를 텍스트로 변환하는 절차이다. 본 시스템은 실시간 디지털 사이니지 환경에서의 운용을 목표로 하기 때문에, 단순한 인식 정확도뿐만 아니라 응답 속도와 자원 효율성을 핵심 설계 기준으로 두었다.

STT 엔진은 OpenAI의 Whisper 모델을 고속 추론에 적합하도록 최적화한 faster-whisper 라이브러리를 사용하였다. 디지털 사이니지의 제한된 연산 자원(CPU/On-device AI)을 고려해, 모델 가중치는 INT8 양자화(Quantization)를 적용하였다. 이는 소형 장치에서의 실시간 음성 향상 모델 연구[15]에서 중요하게 다루는 핵심 기법이며, FP32(32-bit Floating Point) 대비 약 4배의 메모리 절감과 추론 속도 향상을 제공하면서도 한국어 인식 성능(WER) 저하는 최소화한다. 실시간성을 확보를 위해 모델 규모는 small 버전을 채택하였다. STT 프로세스는 영상 처리 지연을 방지하기 위해 별도의 비동기 스레드에서 실행되며, 오디오 데이터는 BLOCK_SIZE(1024 샘플) 단위로 수집되어 누적 버퍼에 저장된다. 이후 2.0초(TRANSCRIPTION_INTERNAL) 주기로 텍스트 변환이 수행된다.

이 단계에서 핵심은 시각적 게이팅(Visual Gating)의 최종 적용이다. 1단계에서 얻은 person_detected와 is_speaking 플래그를 STT 스레드와 공유해, 화면에 사람이 없거나 발화가 없으면 오디오 버퍼를 즉시 폐기(Drop)한다. 다만 문장 말미가 잘리는 문제를 막기 위해 발화 종료 후 1.5초의 유예 시간(PAUSE_THRESHOLD)을 유지한다. Whisper는 입력이 침묵·잡음일 때 학습 데이터의 상투적 문구(예: “MBC 뉴스”, “구독과 좋아요”, 등)를 생성하는 환각(Hallucination) 문제가 있다. 공공장소 사이니지 환경에서는 치명적이므로 본 시스템은 다음의 이중 필터링을 적용한다.

(1) 입력 오디오의 RMS 에너지가 0.001 미만이면 추론을 건너뛴다.

(2) 인식 결과가 짧고(15자 미만) 사전 정의한 환각 키워드를 포함하면 즉시 출력을 무시한다.

4. 실험결과 및 성능분석

4.1 실험 환경 및 평가 지표

본 연구에서 제안한 시스템은 실시간 엣지 환경에서 동작하도록 설계되었지만, 성능 평가 서는 실험의 재현성과 공정성을 확보하기 위해 사전에 녹화된 영상을 활용했다. 실험에는 약 30초 분량의 바람 소리와 도로 소음이 섞인 야외 촬영 영상을 사용하였고, 일반적인 노트북(Windows 11 Pro, Intel Core i5-11300H, 16GB RAM, NVIDIA GeForce RTX 3050) 환경에서 진행되었다.

성능 평가는 시스템의 효율과 신호 품질을 종합적으로 살펴보기 위해 RTF(Real Time Factor), NR(Noise Reduction), Speech Loss 세 가지 정량적 지표를 사용했다. 실제 환경에서 수집하는 데이터 특성상 정답 스크립트를 확보하기 어렵기 때문에 CER(문자 오류율)이나 WER(단어 오류율)과 같은 정확도 지표를 적용하는 대신, 시스템이 실사용 상황에서 얼마나 효과적인지를 검증하는 데 초점을 맞추었다. 이를 위해 표 4와 같이 원본(Raw), 일반적인 잡음 제거(Standard), 그리고 제안 모델(Proposed) 세 가지 실험군을 비교 분석하였다.

표 4. 3가지 실험군
Table 4. Three experimental groups

실험군(Case)	설명	예상 결과
Raw	원본 오디오	STT 인식률 최하(기준점)
Standard Noise Cancellation	상시 잡음 제거	잡음저하, 음성왜곡 상승 → STT 성능 저하·CPU 부하
Proposed	선택적 잡음 제거	음성 왜곡 최소·STT 정확도, ·연산 효율 향상

4.2 실험 결과

제안하는 시스템(Case 3)의 성능은 표 5에서 보여준다. RTF은 0.134로 나타났다. 1초짜리 오디오를 처리하는 데 0.134초밖에 걸리지 않음을 뜻하며, 입술 인식(Visual VAD)과 잡음 제거까지 모두 수행했음에도 일반적인 잡음 제거 방식(Standard, RTF 0.111)과 비슷한 수준의 속도를 유지했다. 이는 저사양 환경에서도 충분히 실시간 처리가 가능하다는 것을 보여준다. 잡음 제거 성능(NR)에서도 제안한 모델은 15.67dB의 감쇠량을 기록해, Standard 방식(8.20dB) 보다 약 1.9배 더 강하게 소음을 억제하는 성능을 보였다. 특히 그림 3의 제안 모델의 Speech Loss(14.59dB) 수치가 높게 나타난 것은 음성 정보가 손상됐다는 의미가 아니라, 발화 구간에 섞여있던 바람 소리 같은 배경 잡음이 효과적으로 제거되면서 전체 신호 에너지가 줄어든 결과로, 오히려 긍정적으로 볼 수 있다.

표 5. 실험군 성능 비교
Table 5. Performance comparison of experimental group

구분	RTF	NR	Speech Loss
Raw	0.00	-	-
Standard Noise Cancellation	0.111	8.20	6.97
Proposed	0.134	15.67	14.59

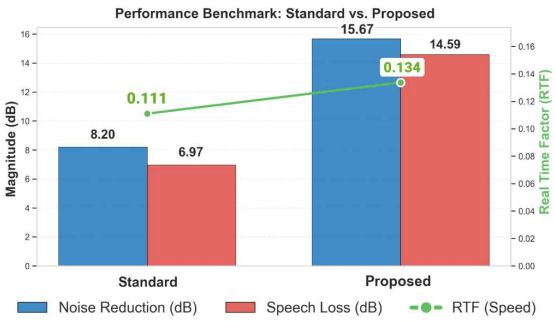

그림 3. 성능 결과 비교 (RTF, Speech Loss 및 NR)
Fig. 3. Performance Comparison (RTF, Speech Loss, and NR)

5. 결 론

본 논문은 공공장소의 불확실한 소음환경 내 기존 에너지 기반 음성 VAD가 갖는 한계를 극복하고, 저사양 엣지 환경에서의 실용적인 구동 효율성을 최우선으로 하는 시각 정보 기반의 멀티모달 전처리 파이프라인을 제안하였다. 제안하는 시스템은 사용자의 입술 움직임(LAR)을 트리거로 활용하여 발화 구간을 명확히 판별, 타임스탬프 패딩(Padding) 로직과 시각적 판별의 확신도를 나타내는 Confidence 기반의 적응형(Adaptive) 후처리를 통해 데이터 유실을 방지하고 안정성을 확보하였다. 기술적 측면에서 본 시스템은 '연산 효율성'과 '저지연 확보'에 집중하여 비발화 구간의 무거운 오디오 처리 및 STT 연산을 원천 차단하는 게이팅구조, 스트리밍 환경에 최적화된 상태 보존형(Stateful) 청크 단위 처리 방식을 채택하였다. 이를 통해 고성능 GPU 서버가 부재한 저전력 임베디드 환경에서도 즉각적 응답 속도를 유지하는 것이 가능하다. RTF는 0.134를 기록하여, 일반적인 잡음제거 파이프라인 대비 연산 과정이 추가됐음에도 불구하고 RTF를 비슷한 수준으로 유지하였다. NR은 15.67dB를 달성하여 기존 방법대비 잡음제거 성능이 개선되었음을 보인다. Speech Loss는 14.59dB로 나타나 배경 잡음이 효과적으로 제거된 결과로 해석할 수 있다.

제안한 연구는 디지털 기기 조작에 미숙하거나 터치 인터페이스 사용에 적응하지 못하는 고령층 및 디지털 소외 계층의 기술 접근 장벽을 낮추고 기존 터치 인터페이스의 한계를 극복함을 통해 세대를 불문한 포용적 디지털 환경 조성은 기술 접근 장벽으로 인해 억제되었던 특정층의 경제 활동을 촉진할 수 있을 것이며, 소비 활동을 격려하는 데에 이바지할 수 있기를 기대한다.

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

참 고 문 헌

[1] J. Park, W. Kim, D. K. Han, H. Ko, "Voice Activity Detection in Noisy Environments Based on Double-Combined Fourier Transform and Line Fitting", The Scientific World Journal, vol. 2014, 146040, Aug 2014. DOI: 10.1155/2014/146040

[2] T. A. Ma, S. Yin, L.-C. Yang, S. Zhang, "Real-Time Audio-Visual Speech Enhancement Using Pre-trained Visual Representations", Interspeech 2025, arXiv:2507.21448 [eess.AS], July 2025.

[3] C. Yu, J. Yu, Z. Qian, Y. Tan, "Improvement of Acoustic Models Fused with Lip Visual Information for Low-Resource Speech", Sensors, vol. 23, no. 4, 2071, Feb 2023. DOI: 10.3390/s23042071

[4] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 27, no. 2, pp. 113-120, 1979. DOI: 10.1109/TASSP.1979.1163209

[5] P. C. Loizou, Speech Enhancement: Theory and Practice, 2nd ed., CRC Press, 2013. ISBN: 978-1466504219

[6] J. Chen and X. Ran, "Deep Learning With Edge Computing: A Review", Proceedings of the IEEE, vol. 107, no. 8, pp. 1655-1674, 2019. DOI: 10.1109/JPROC.2019.2921977

[7] H. McGurk and J. MacDonald, "Hearing lips and seeing voices", Nature, vol. 264,

pp. 746-748, 1976. DOI: 10.1038/264746a0
[8] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Deep Audio-Visual Speech Recognition”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018. DOI: 10.1109/TPAMI.2018.2889052
[9] A. Ephrat et al., “Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation”, ACM Transactions on Graphics (SIGGRAPH), vol. 37, no. 4, 2018. DOI: 10.1145/3197517.3201357
[10] J. Yu et al., “Audio-Visual Recognition of Overlapped Speech for the LRS2 Dataset”, ICASSP 2020, 2020. DOI: 10.1109/ICASSP40206.2020.9054429
[11] G. Jocher et al., “YOLOv5 by Ultralytics”, 2020. [Online]. DOI: 10.5281/zenodo.3908559
[12] Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann, “Real-time Facial Surface Geometry from Monocular Video on Mobile GPUs”, in Proc. CVPR Workshop on Computer Vision for Augmented and Virtual Reality, Long Beach, CA, USA, Jun. 2019, pp. 1-5. DOI: 10.1109/CVPRW.2019.00236
[13] M. Satyanarayanan, “The Emergence of Edge Computing”, IEEE Computer, vol. 50, no. 1, pp. 30-39, 2017. DOI: 10.1109/MC.2017.9
[14] D. K. Ha, S. J. Cho, K. K. Jin, and O. K. Shin, “Voice Activity Detection in Noisy Environments Based on Entropy Difference and Signal Energy”, Journal of Advanced Marine Engineering and Technology, vol. 32, no. 5, pp. 768-774, 2008. DOI: 10.5916/jkosme.2008.32.5.768
[15] S. Y. Lee, H. S. Kim, and J. W. Shin, “Develop Deep Learning-based Real-time Speech Enhancement models for Small Devices”, in Proc. Korea Institute of Broadcast and Media Engineers (KIBME) Autumn Conference, 2024, pp. 123-125. https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE11995439

저 자 소 개

고현지(Hyeon-Ji Ko)

2023~현재 신한대학교 소프트웨어융합학과 재학
<주관심분야> Cybersecurity, Network Systems, Artifical Intelligence

서원후(Won-Hu Seo)

2023~현재 신한대학교 소프트웨어융합학과 재학
<주관심분야> Database Management, Network Security

최용수(YongSoo CHOI)

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