한국음향학회지 제44권 제5호 pp. 533~539 (2025)
*The Journal of the Acoustical Society of Korea* Vol.44, No.5 (2025)
https://doi.org/10.7776/ASK.2025.44.5.533

pISSN : 1225-4428
eISSN : 2287-3775

# 음성 변형 추적 : 강건한 음성 처리를 위한 노이즈 분류 접근법

# Speech distortion tracing : A noise classification approach for robust speech processing

이웅재,[1] 정수환[1†]

(Woongjae Lee[1] and Souhwan Jung[1†])

[1]숭실대학교 정보통신공학과

(Received August 1, 2025; revised August 28, 2025; accepted September 1, 2025)

**초 록**: 다양한 노이즈 유형을 이해하고 분류하는 것은 음성 처리 모델의 강건성을 향상시키는 데 있어 매우 중요하다. 기존 연구들은 주로 백색 잡음이나 배경 소음과 같은 환경 노이즈에 초점을 맞추었으나, 피치 시프트, 타임 스트레치와 같은 음성 변조의 영향은 상대적으로 덜 탐구되어 왔다. 본 연구는 이러한 한계를 극복하기 위해 음성 변형을 포함한 다양한 노이즈 유형을 효과적으로 구분할 수 있는 노이즈 분류 모델을 제안한다. LibriSpeech, Voice Cloning ToolKit (VCTK), DSD-Corpus, TIMIT 데이터셋을 활용하여 538,000개의 노이즈 데이터셋을 구축하였으며, Spectrogram, MFCC, F0를 융합한 Multi-Feature Fusion 모델을 설계하였다. 제안된 모델은 95 %의 정확도를 달성하였으며, 7.2 M 파라미터로 실시간 전처리에 적합한 경량성을 확보하였다. 이는 노이즈 분류 기반 적응형 전처리가 다양한 노이즈 환경에서의 음성 처리 모델 강건성 향상에 기여할 수 있음을 시사한다.

**핵심용어**: 노이즈 분류, 음성 변조, 음성 처리, 오디오 딥페이크 탐지, 자동 화자 검증

**ABSTRACT:** Understanding and classifying various types of noise is crucial for improving the robustness of speech processing models. While previous studies have primarily focused on environmental noise such as white noise and background noise, the effects of speech manipulation including pitch shift and time stretch have been relatively less explored. This study proposes a noise classification model that can effectively distinguish various types of noise including speech manipulation to overcome these limitations. We constructed a noise dataset of 538,000 samples using LibriSpeech, Voice Cloning ToolKit (VCTK), DSD-Corpus, and TIMIT datasets, and designed a Multi-Feature Fusion model that integrates spectrogram, MFCC, and F0 features. The proposed model achieved 95 % accuracy with 7.2 M parameters, demonstrating suitable lightweight characteristics for real-time preprocessing. This suggests that noise classification-based adaptive preprocessing can effectively contribute to enhancing the robustness of speech processing models in various noise environments.

**Keywords:** Noise classification, Speech manipulation, Speech processing, Audio deepfake detection, Automatic speaker verification

**PACS numbers:** 43.60.Bf, 43.60.Mn, 43.60.Lq

## I. 서 론

최근 음성 처리 기술의 발전, 특히 자동 음성 인식 (Automatic Speech Recognition, ASR), 자동 화자 검증 (Automatic Speaker Verification, ASV), 오디오 딥페이크 탐지(Audio Deepfake Detection, ADD) 모델의 성능 향상은 인증 및 검증 시스템의 신뢰도를 크게 높여 왔다. 그러나 이러한 모델들은 주로 깨끗한 환경에서 개발 및 평가되기 때문에, 실제 환경에서 마주치는 다양한 노이즈 요인에 취약하다는 한계가 있다.[1]

† **Corresponding author:** Souhwan Jung (souhwanj@ssu.ac.kr)
Department of Electronic Engineering, Soongsil University,46, Sadang-ro, Dongjak-gu, Seoul 06978, Republic of Korea
(Tel: 82-2-826-9197)

Copyrightⓒ 2025 The Acoustical Society of Korea. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

특히, 음성 변조는 기존의 환경 노이즈와 달리 음성의 신호 스펙트럼과 시간적 특성을 동시에 변화시키면서도 청취자에게는 자연스러운 음성처럼 인식될 수 있는 점이 특징이다. 예를 들어 피치 시프트, 타임 스트레치, 오토튠과 같은 변형은 음성 신호를 인위적으로 왜곡하지만 명료도가 유지되면서 정상 음성으로 받아들여질 수 있다. 이런 음성 변조는 음성 처리 모델의 혼동을 더욱 가중시켜, 전통적인 환경 노이즈보다 더 심각한 성능 저하를 유발하는 것으로 나타났다.

이와 같은 문제를 극복하기 위해, 본 연구는 음성 변조를 포함한 다양한 노이즈 유형이 음성 처리 모델의 성능에 미치는 영향을 체계적으로 분석한다. 이러한 복합적 노이즈 환경에서도 모델의 강건성을 실직적으로 향상시킬 수 잇는 노이즈 분류 전략을 제안하고, 그 필요성을 강조한다.

## II. 선행 연구

선행 연구에서는 합성 노이즈가 음성 향상과 같은 음성 처리 작업에 미치는 영향을 다양하게 분석하였다.[2] 특히 노이즈 강건성을 향상시키기 위한 데이터 증강 기법들이 활발히 연구되어 왔으며,[3] 주파수 마스킹, 시간 왜곡 등 다양한 증강 방법과 Attention 메커니즘, CNN 기반 모델의 결합을 통해 소음 환경에서의 음성 감정 분류 성능을 향상시키는 사례도 연구되었다.[4]

자동 음성 인식(ASR) 분야에서는 스펙트럼 기반의 endpoint detection 기법을 활용하여 노이즈가 포함된 환경에서도 정확한 음성 분할을 수행하려는 시도가 이루어졌다.[5] 더불어, 다양한 노이즈 유형에 강건한 분류 성능을 확보하기 위해 다중 임베딩 기법이 제안되었으며, 이는 서로 다른 음향 특징들의 융합을 통해 분류 정확도를 크게 향상시키는 것으로 나타났다.[6]

그러나 이러한 기존 연구들은 주로 배경 소음 및 합성 노이즈에 집중한 반면, 음성 변조를 독립적인 노이즈 유형으로 인지하고 분류하려는 시도는 이루어지지 않았다. 음성 변조는 신호의 스펙트럼과 시간적 특성을 동시에 변화시키면서도 자연스러운 청취감을 유지한다는 점에서, 전통적인 환경 노이즈와 구분되는 별도의 연구 접근이 필요하다. 따라서 본 연구는 음성 변조를 포함한 다양한 노이즈 유형을 효과적으로 분류할 수 있는 새로운 노이즈 분류 모델 설계 및 평가를 통해 이 연구 격차를 해소하는 것을 목표로 한다.

## III. 노이즈 데이터셋 구축

음성 처리 모델의 노이즈 강건성을 평가하고, 노이즈 분류 모델을 훈련시키기 위해, 본 연구에서는 다양한 노이즈 데이터셋을 구축하였다. LibriSpeech,[7] Voice Cloning ToolKit(VCTK),[8] DSD-Corpus[9] 그리고 TIMIT[10]을 음성 데이터 소스로 활용하였다. 4가지 데이터셋 모두 다양한 화자를 포함하며, 특히 DSD-Corpus는 Artificial Intelligence(AI) 기반 합성 음성도 포함함으로써, 실제 음성과 합성 음성 모두에서의 노이즈 효과를 포괄적으로 분석할 수 있게 한다.

구축한 데이터셋의 노이즈 샘플은 실제 환경을 반영하고자 배경 소음, 합성 노이즈, 음성 변조의 세 가

Table 1. Methods for generating noise and speech manipulation data.

| |
|---|
| Clean : Original, unaltered speech samples from LibriSpeech, VCTK, TIMIT, DSD-Corpus |
| Add Background Music : Augmented using the MUSAN[11] dataset with an SNR range of 10 dB ~ 20 dB |
| Add Background Noise : Augmented with environmental and natural noise from the ESC-50[12] dataset within an SNR range of −6 dB to 3 dB |
| Overlapping Speech : Augmented with speech from other speakers in the MUSAN dataset at an SNR range of 10 dB - 20 dB |
| White Noise : Mixed with random noise sampled from a Gaussian distribution (mean = 0, variance = 1) at an SNR range of −10 dB to 10 dB |
| Pink Noise : Mixed with FFT-based spectrally scaled pink noise at an SNR range of −10 dB to 10 dB |
| Pitch Shift : Adjusted by ±2 to ±5 semitones using the librosa.effects.pitch_shift() function |
| Time Stretch : Adjusted to 0.7x/1.7x speed using the librosa.effects.time_stretch() function |
| Auto Tune : Pitch was corrected to the nearest note in a predefined musical scale using an open-source implementation[13] |
| Reverberation : Apply realistic reverberation using the fftconvolve function with the RIR[14] dataset |

지 주요 유형으로 구분하였다. 각 노이즈 유형은 실제 상황을 모사함과 동시에 언어 정보의 보존을 최대한 고려하여 설계되었으며, Table 1에 명시된 총 10개의 클래스를 포함하여 다양한 노이즈 환경을 포괄한다.

배경 소음의 경우, 선행 연구[15]를 참고하여 신호대 잡음비(Signal to Noise Ratio, SNR)를 −6 dB에서 3 dB 범위로 설정하여 실제 환경을 시뮬레이션하였다. 합성 노이즈는 백색 노이즈와 핑크 노이즈로 구성하였으며, 디지털 신호 처리 기법을 활용해 SNR [−10, 10] 사이의 범위로 합성하였다. 음성 변조 유형으로는 피치 시프트, 타임 스트레치, 오토튠이 포함되며, 특히 피치 시프트는 ±5 반음 이내로 조절하여 인위적인 왜곡(Chipmunk effect)[16]을 방지하였다. 변조 강도는 언어 정보가 최대한 유지될 수 있도록 세밀하게 조정되었다.

데이터셋은 학습 및 개발용으로 LibriSpeech, VCTK, DSD-Corpus를 사용하였고, 평가용 데이터셋은 TIMIT을 별도로 배정하여 노이즈 분류 모델의 일반화 성능을 검증하였다. 총 538,000개의 오디오 샘플 중 270,269개는 학습용, 115,831개는 개발용, 152,900개는 평가용으로 분할되었으며, 모든 데이터셋은 화자 단위로 중복 없이 구성되었다. 특히 TIMIT 기반 평가 데이터는 학습 및 개발 데이터와 전혀 다른 문장 구조와 발화 스타일을 포함하도록 설계되어, 모델이 음성 내용이나 문장 스타일의 영향을 배제하고 오직 노이즈 특성 자체만을 정확히 분류하는지를 객관적으로 평가할 수 있도록 하였다.

## IV. 노이즈 유형별 음성 처리 모델 성능 영향 평가

본 섹션에서는, 다양한 노이즈 유형이 자동 화자 검증 모델과 오디오 딥페이크 탐지 모델의 성능에 미치는 영향을 분석하기 위해 LibriSpeech와 VCTK를 기반으로 한 데이터셋을 활용하였다. ASV 모델 평가는 LibriSpeech와 VCTK의 음성 데이터를 통합하여 진행하였으며, ADD 모델의 경우에는 학습이 VCTK 기반의 ASVSpoof19 데이터셋[17]으로 구성됨을 고려하여 In-Domain(VCTK) 및 Out-of-Domain(LibriSpeech)으로 구분하여 실험을 설계하였다. 이처럼 도메인 기반 구분은, 모델이 훈련에 사용된 데이터와 동일한 환경에서는 상대적으로 높은 성능을 보일 수 있으나, 이전에 경험하지 못한 새로운 도메인에서는 일반화 성능이 급격히 저하될 수 있다는 점을 정량적으로 검증하기 위해서이다. 따라서, 각 도메인 조건에서의 노이즈 민감도 및 성능 편차를 체계적으로 분석하고자 하였다.

실험에 사용된 ASV 모델은 ECAPA-TDNN,[18] NeXt-TDNN,[19] ECAPA2[20]로 구성하였으며, ADD 모델은 AASIST,[21] AASIST-SSL,[22] Conformer-TCM[23]을 활용하였다. ASV 모델의 성능 평가는 Equal Error Rate (EER)을 기준으로 산출하였다. EER은 거짓 수락률(False Acceptance Rate, FAR)과 거짓 거부율(False Rejection Rate, FRR)이 동일해지는 시점의 오류율로, 바이오인식 분야에서 널리 사용되는 대표적인 평가지표이다. 한편, ADD 모델은 본 연구의 초점이 다양한 노이즈 조건에서 bonafide 음성 탐지의 신뢰성을 평가하는 데 있으므로, spoof 음성을 제외한 bonafide 음성에 대한 탐지 정확도를 성능지표로 채택하였다.

이와 같은 평가 지표의 선택은 실제 환경에서 발생할 수 있는 다양한 노이즈에 노출된 음성에 대해 시스템이 얼마나 견고하게 화자를 판별하고 신뢰할 수 있는 음성임을 판정할 수 있는지를 정밀하게 측정하기 위함이다.

Table 2의 결과에 따르면 ADD 모델들은 전반적으로 깨끗한 환경에서는 높은 정확도를 보였으나, 노이즈 유형에 따라 성능 편차가 뚜렷하게 나타났다. 특히 In-Domain 환경(VCTK)에서도 일부 노이즈 조건에 서는 정확도가 급격히 하락하는 양상이 관찰되었으며, Out-of-Domain 환경(LibriSpeech)에서는 이러한 성능 저하가 더욱 심화되었다.

실 생활에서 흔히 접할 수 있는 노이즈(주변 소음, 음악, 다중 화자의 발화 등)는 대부분의 모델에서 탐지 성능 저하를 유발했다. 그 중에서도 백색 잡음과 같이 전 주파수 대역에 걸쳐 에너지가 고르게 분포된 노이즈는 음성 주요 특징을 마스킹하여 탐지 정확도를 크게 저하시키는 경향을 보였다. 음성 변조 유형에 대해서는 모델간 성능 편차가 특히 두드러지게 나타났다. 피치 시프트, 타임 스트레치, 오토튠

Table 2. Accuracy of ADD models under Various Noise Conditions in In-Domain (VCTK) and Out-of-Domain (LibriSpeech) Scenarios.

| | VCTK (In-Domin) | | | LibrSpeech (Out-of-Domain) | | |
|---|---|---|---|---|---|---|
| Accuracy (%) | AASIST | SSL-AASIST | ConformerTCM | AASIST | SSL-AASIST | ConformerTCM |
| Clean | 99.78 | 97.85 | 99.52 | 85.90 | 77.79 | 74.64 |
| Background noise | 13.46 | 64.29 | 58.34 | 6.46 | 37.8 | 24.72 |
| Background music | 93.34 | 93.77 | 93.05 | 67.32 | 68.42 | 58.97 |
| Overlapping Speech | 94.62 | 73.06 | 61.26 | 62.31 | 48.39 | 31.11 |
| White noise | 13.85 | 49.88 | 79.76 | 3.13 | 18.63 | 34.83 |
| Pink noise | 72.44 | 95.87 | 96.94 | 51.84 | 63.01 | 63.45 |
| Pitch Shift | 23.90 | 4.07 | 10.82 | 15.19 | 3.02 | 4.87 |
| Time Stretch | 21.61 | 3.28 | 8.25 | 13.33 | 0.33 | 1.73 |
| Auto Tune | 98.13 | 15.86 | 36.93 | 70.11 | 3.6 | 6.69 |
| Reverberation | 22.46 | 34.29 | 71.14 | 33.25 | 15.53 | 19.55 |

Table 3. EER of ASV models under various noise conditions.

| EER (%) | ECAPA | ECAPA2 | NeXt-TDNN |
|---|---|---|---|
| Clean | 0.48 | 1.09 | 1.33 |
| Background noise | 4.85 | 3.76 | 5.58 |
| Background music | 1.00 | 0.82 | 1.73 |
| Overlapping Speech | 3.03 | 1.70 | 4.48 |
| White noise | 7.39 | 6.79 | 8.97 |
| Pink noise | 1.09 | 1.09 | 2.55 |
| Pitch Shift | 42.67 | 26.91 | 44.73 |
| Time Stretch | 2.62 | 3.49 | 6.04 |
| Auto Tune | 0.73 | 1.21 | 1.94 |
| Reverberation | 2.67 | 2.91 | 3.64 |
| Pooled EER | 6.63 | 5.90 | 8.13 |

과 같은 조작은 음성의 주파수 및 시간 구조를 동시에 변형하며, 단순한 환경 노이즈보다도 더 심각한 성능 저하를 유발하였다. 특히 타임 스트레치나 오토튠의 경우, 일부 모델은 탐지를 거의 수행하지 못하는 수준의 결과를 보이며, 구조적으로 해당 변조 유형에 취약함을 보여주었다.

Table 3의 결과를 보면 ASV 모델은 전반적으로 깨끗한 환경에서는 매우 낮은 오류율을 보이며 안정적인 성능을 나타냈으나, 노이즈 유형에 따라 뚜렷한 성능 편차가 나타났다. 특히 피치 시프트 조건에서는 모든 모델에서 가장 심각한 오류율 상승이 관찰되었다. 이는 피치 시프트가 화자의 고유한 주파수 패턴과 음성 특징을 변형하여 화자 인식을 위한 필수 정보를 손상시키기 때문으로 분석된다. 반면, pink noise나 배경 음악과 같은 조건에서는 상대적으로 낮은 오류율을 유지하여, 이러한 환경에서는 모델이 상대적으로 강건함을 보여주었다.

이러한 실험 결과는 실제 환경에서 흔히 접할 수 있는 다양한 노이즈와 음성 변조가 기존 음성 처리 모델의 성능 저하를 유발하며, 특히 모델이 학습된 도메인에서 벗어난 환경에서는 그 영향이 더욱 심화될 수 있음을 시사한다. 따라서 본 연구에서 제안하는 바와 같이, 입력 음성에 포함된 노이즈 유형을 사전에 분류하여 각 노이즈 환경에 적합한 전처리 및 후처리 기법을 적응적으로 적용할 수 있는 시스템의 필요성이 강조된다.

## V. 적응형 전처리를 위한 노이즈 분류 모델 설계

본 연구에서는 다양한 노이즈 유형을 효과적으로 분류하기 위해 여러 신경망 구조와 입력 특징 조합을 적용한 노이즈 분류 모델을 설계하고, 이들의 성능을 비교 평가하였다. 제안된 노이즈 분류 모델은 단순히 정확도 향상에 그치지 않고, 실제 음성 처리 시스템의 전처리 단계에 적용 가능한 실용성과 다양한 환경에서의 일반화 성능 확보에 중점을 두고 설계되었다.

모델 구성은 다음과 같다:

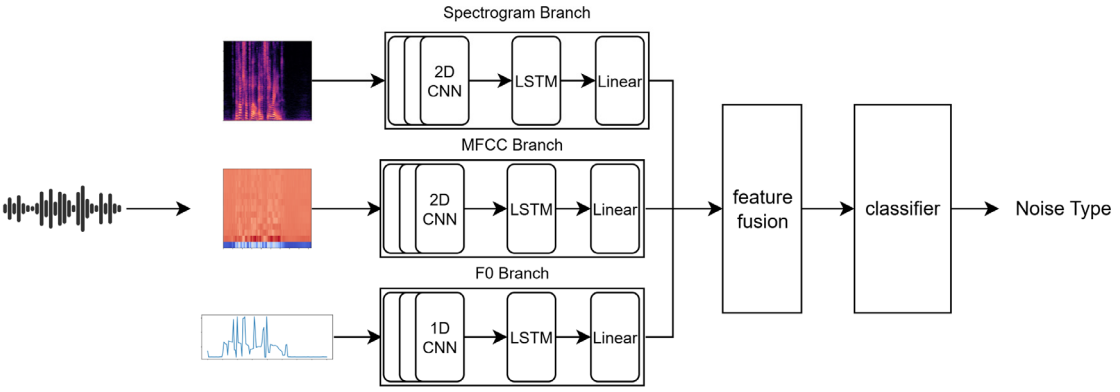



Fig. 1. (Color available online) Architecture of the proposed multi-feature fusion model.

- Wav2Vec 2.0 + linear : Wav2Vec 2.0[24]는 대규모 음성 데이터를 대상으로 비지도 학습이 된 모델로, 원시 파형에서 강력한 임베딩 표현을 추출한다. 이러한 특성은 다양한 노이즈 환경에서도 견고한 특성 표현력을 제공할 것으로 기대되어 선택하였다.
- SSAST[25] : 트랜스포머 기반의 self-attention 메커니즘을 활용하여 긴 시계열 정보와 복잡한 스펙트럼 구조를 효율적으로 모델링한다. Wav2Vec 2.0 모델과는 상호 보완적으로, 원시 파형에서 포착하기 어려운 미세한 세부적인 스펙트럼 특성에 집중하여 보다 정밀한 노이즈 분류가 가능하다.
- CNN + LSTM[26] : CNN을 통해 음성 신호 내의 지역적인 패턴을 추출하고, LSTM으로 시간적 변화 및 연속성을 반영하는 하이브리드 구조이다. 특히, 시간-주파수 영역에서 변동성이 큰 노이즈 유형에 대해 효과적으로 대응할 수 있도록 설계되었다.
- Multi-feature fusion : 노이즈 유형별로 민감하게 반응하는 음향 특징이 다를 수 있다는 점에 착안하여 Fig. 1과 같이 설계되었다. Spectrogram, MFCC, F0의 서로 다른 특성을 지닌 세 가지 음향 특징을 각각 CNN + LSTM 구조로 처리하여 시간적 연속성과 주파수적 변동성을 동시에 반영하였다. 세 브랜치 모두 동일한 구조적 효과를 지니며, 각 브랜치 별로 1024차원의 임베딩을 산출한다. 이렇게 얻어진 세 가지 임베딩은 벡터 연결 방식을 통해 융합되며, 이후 분류기는 두 개의 Linear 계층으로 구성된다. 첫 번째 계층에서는 고차원의 정보를 압축하고, 두 번째 계층은 최종 클래스 유형으로 매핑한다. Dropout이 중간에 삽입되어 과적합을 방지하고 일반화 성능을 높였다.

Table 4. Evaluation results of noise classification models.

| | Feature | Parameter | Evaluation |
|---|---|---|---|
| Wav2Vec2 + Linear | Represented vector | 319.6 M | 91 % |
| SSAST - base | Spectrogram | 87 M | 98.5 % |
| SSAST - small | Spectrogram | 22.6 M | 97.5 % |
| CNN + LSTM | MFCC | 0.3 M | 80 % |
| Multfi feature fusion | Spectrogram + MFCC + F0 | 7.2 M | 95 % |

모든 모델은 앞에서 구축한 노이즈 데이터셋을 활용하여 훈련되었으며, 평가에는 훈련 단계에 포함되지 않은 TIMIT 데이터셋과 DSD-Corpus를 사용하여, 문맥적 및 화자 영향을 최소화함과 동시에 노이즈 유형별 분류 성능을 객관적으로 비교하였다.

Table 4에서는 각 노이즈 분류 모델의 성능 평가 결과를 확인할 수 있다. 전반적으로 모든 모델이 우수한 성능을 보였으나, 모델 구조와 입력 특징에 따라 차이가 존재하였다. 특히 SSAST 모델의 경우 주파수 특성을 효과적으로 학습함으로써 가장 뛰어난 노이즈 분류 성능을 보였다. 경량화된 SSAST-small 모델 역시 높은 정확도를 유지하여, 제한된 자원을 갖춘 시스템 환경에서도 활용 가능함을 확인하였다. 또한, Multi-feature fusion 모델은 다양한 음향 특

징을 융합하여 단일 특징을 사용하는 CNN + LSTM 모델보다 크게 향상된 성능을 보였다. 이는 노이즈 유형마다 서로 다른 음향적 특성에 민감하게 대응할 수 있는 다중 특징 융합 전략이 노이즈 분류에 매우 효과적임을 시사한다. 반면 Wav2Vec 2.0 + Linear 모델은 원시 파형에서 추출한 representation vector를 활용하여 비교적 우수한 성능을 보였지만, 대규모의 파라미터 수와 waveform의 입력 의존성으로 인해 특정 복잡한 노이즈 조건에서는 한계가 관찰되었다.

종합적으로 실시간 음성 처리 시스템의 전처리 단계에서 노이즈 유형을 정확하게 식별하고 적절히 대응하기 위해서는 모델이 높은 정확도와 적은 파라미터 수를 동시에 만족해야 한다. 본 연구 결과에 따르면 Multi-Feature Fusion 모델이 이러한 요구사항을 균형있게 충족함으로써 실제 응용에 가장 적합한 모델임을 확인할 수 있었다.

## VI. 결 론

본 연구에서는 음성 처리 시스템의 강건성을 향상시키기 위한 전처리 단계에서의 노이즈 분류 모델의 필요성을 제안하고, 이를 다양한 실험을 통해 검증하였다. 이를 위해 실제 환경을 반영한 다중 노이즈 유형의 대규모 데이터셋을 구축하였으며, 맥락적 요소에 의존하지 않는 노이즈 분류 성능을 체계적으로 평가하였다. 실험 결과 기존 음성 처리 모델들은 실제 환경에서 흔히 발생하는 다양한 노이즈 및 음성 변조 조건에서 심각한 성능 저하를 겪었으며, 이러한 문제를 해결하기 위해 노이즈 유형별 분류 및 이에 기반한 적응적 전처리 기법이 필수적임을 확인하였다. 특히, Multi-Feature Fusion 노이즈 분류 모델은 적은 파라미터 수 대비 탁월한 분류 성능을 보여 실시간 음성 처리 시스템의 전처리 단계에 적용할 경우 높은 실용성과 강건성을 동시에 제공할 수 있을 것으로 기대된다.

## VII. 향후 연구 방향

향후 연구에서는 본 연구에서 검증한 노이즈 분류 기반 전처리 모델을 실제 음성 처리 시스템과 통합하는 다양한 방안을 모색할 계획이다.

첫째, Low-Rank Adaptation(LoRA) 기반 어댑터를 활용하여 각 노이즈 타입에 특화된 LoRA 모듈을 사전에 학습해 두고, 입력 신호의 노이즈 유형에 따라 해당 LoRA를 백엔드 모델에 동적으로 라우팅하는 방식을 고려한다. 이 접근법은 노이즈별로 백엔드 전체 모델을 새롭게 파인튜닝할 필요 없이, 경량화된 어댑터만 적용함으로써 자원 효율성과 환경 적응성을 동시에 확보할 수 있다는 장점이 있다. 이를 통해 실시간 환경에서도 노이즈별 최적화된 탐지 및 처리가 가능할 것으로 기대된다.

둘째, 노이즈 유형별로 특화된 음성 향상 기법을 적용하는 자가적응 보상 방식을 도입하여, 다양한 실제 환경에서 모델의 강건성과 적응력을 한층 강화할 수 있을 것으로 기대된다.

## 감사의 글

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학ICT연구센터(ITRC)의 지원(IITP-2025-RS-2020-II201602, 50 %)과 과학기술정보통신부 및 정보통신기획평가원의 융합보안핵심인재양성사업의 연구 결과로 수행되었음 (IITP-2025-RS-2024-00426853, 50 %).

## References

1. S. Katkov, A. Liotta, and A. Vietti, “Evaluating the robustness of ASR systems in adverse acoustic conditions,” Proc. 5th IDSTA, 76-80 (2024).
2. V. K. Singh, K. Sharma, and S. N. Sur, “A survey on preprocessing and classification techniques for acoustic scene,” Expert Syst. Appl. **229**, 120520 (2023).
3. T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, “Audio augmentation for speech recognition,” Proc. Interspeech, 3586-3590 (2015).
4. L. Wijayasingha and J. A. Stankovic, “Robustness to noise for speech emotion classification using CNNs and attention mechanisms,” Smart Health, **19**, 100165 (2021).
5. S. E. Bou-Ghazale and K. Assaleh, “A robust endpoint detection of speech for noisy environments with application to automatic speech recognition,” Proc. IEEE ICASSP, IV-3808 (2002).

6. M. N. Ali, A. Brutti, and D. Falavigna, “Enhancing embeddings for speech classification in noisy conditions,” Proc. Interspeech, 2933-2937 (2022).
7. V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “LibriSpeech: an ASR corpus based on public domain audio books,” Proc. IEEE ICASSP, 5206-5210 (2015).
8. C. Veaux, J. Yamagishi, and K. MacDonald, “CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit,” The Centre for Speech Technology Research (CSTR), University of Edinburgh, Tech. Rep., 2017.
9. T.-P. Doan, H. Dinh-Xuan, T. Ryu, I. Kim, W. Lee, K. Hong, and S. Jung, “Trident of poseidon: A generalized approach for detecting deepfake voices,” Proc. CCS, 2222-2235 (2024).
10. J. S. Garofolo, L. F. Lamel, W. M. Fisher, D. S. Pallett, N. L. Dahlgren, V. Zue, and J. G. Fiscus, “TIMIT acoustic-phonetic continuous speech corpus,” National Institute of Standards and Technology (NIST), Tech. Rep., 1993.
11. D. Snyder, G. Chen, and D. Povey, “MUSAN: A music, speech, and noise corpus,” arXiv preprint arXiv:1510.08484 (2015).
12. K. J. Piczak, “ESC: Dataset for environmental sound classification,” Proc. 23rd ACM Int. Conf. Multimedia, 1015-1018 (2015).
13. *Simple Auto-Tune in Python*, https://github.com/JanWilczek/python-auto-tune, (Last viewed October 10, 2024).
14. T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, “A study on data augmentation of reverberant speech for robust speech recognition,” Proc. IEEE ICASSP, 5220-5224 (2017).
15. J. B. Awotunde, R. O. Ogundokun, F. E. Ayo, and O. E. Matiluko, “Speech segregation in background noise based on deep learning,” IEEE Access **8**, 169568-169575 (2020).
16. S. Rosenzweig, S. Schwär, J. Driedger, and M. Müller, “Adaptive pitch-shifting with applications to intonation adjustment in a cappella recordings,” Proc. 24th DAFx, 121-128 (2021).
17. X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, and Z. H. Ling, “ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech,” Comput. Speech Lang. **64**, 101114 (2020).
18. B. Desplanques, J. Thienpondt, and K. D. Ecapa-tdnn, “Emphasized channel attention, propagation and aggregation in tdnn based speaker verification,” arXiv: 2005.07143 (2020).
19. H. J. Heo, U. H. Shin, R. Lee, Y. Cheon, and H. M. Park, “NeXt-TDNN: Modernizing multi-scale temporal convolution backbone for speaker verification,” Proc. IEEE ICASSP, 11186-11190 (2024).
20. J. Thienpondt and K. Demuynck, “Ecapa2: A hybrid neural network architecture and training strategy for robust speaker embeddings,” Proc. IEEE ASRU, 1-8 (2023).
21. J. W. Jung, H. S. Heo, H. Tak, H. J. Shim, J. S. Chung, B. J. Lee, and N. Evans, “Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks,” Proc. IEEE ICASSP, 6367-6371 (2022).
22. H. Tak, M. Todisco, X. Wang, J. W. Jung, J. Yamagishi, and N. Evans, “Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation,” arXiv:2202.12233 (2022).
23. D. T. Truong, R. Tao, T. Nguyen, H. T. Luong, K. A. Lee, and E. S. Chng, “Temporal-channel modeling in multi-head self-attention for synthetic speech detection,” arXiv:2406.17376 (2024).
24. A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” Proc. NeurIPS, 12449-12460 (2020).
25. Y. Gong, C. I. Lai, Y. A. Chung, and J. Glass, “Ssast: Self-supervised audio spectrogram transformer,” Proc. AAAI Conf. Artif. Intell. 10699-10709 (2022).
26. R. Akter, M. R. Islam, S. K. Debnath, P. K. Sarker, and M. K. Uddin, “A hybrid CNN-LSTM model for environmental sound classification: Leveraging feature engineering and transfer learning,” Digit. Signal Process. **163**, 105234 (2025).

## 저자 약력

▸ 이 웅 재 (Woongjae Lee)



2024년 2월 : 숭실대학교 전자정보공학부 졸업
2024년 9월 ~ 현재 : 숭실대학교 정보통신공학과 석사과정

▸ 정 수 환 (Souhwan Jung)



1985년 2월 : 서울대학교 전자공학과 졸업
1987년 8월 : 서울대학교 전자공학과 석사
1996년 6월 : 미국 Univ. of Washington 박사
1997년 ~ 현재 : 숭실대학교 전자정보공학부 교수
2020년 ~ 현재 : 과학기술정보통신부 ICT 연구센터 AI보안연구센터장