JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 3, pp. 9-19, March 2026
https://doi.org/10.9708/jksci.2026.31.03.009

Adversarial Noise-based Speaker De-identification for Cloud Speech Services

Haram Kang*, Sangwoon Yun**, Jemin Ahn***, Kyungtae Kang****

*M.S. Student, Dept. of Applied Artificial Intelligence, Hanyang University, Seoul, Korea
**Ph.D. Candidate, Dept. of Computer Science and Engineering, Hanyang University, Seoul, Korea
***Postdoctoral Researcher, Research Institute of AI Convergence, Hanyang University, Ansan, Korea
****Professor, Dept. of Artificial Intelligence, Hanyang University, Seoul, Korea

[Abstract]

Cloud-based speech recognition services provide high convenience and accessibility, leading to their widespread adoption across various applications. However, the processing of speech data on remote servers has raised growing concerns about potential privacy breaches stemming from the exposure of speaker information. Although numerous speaker de-identification techniques have been proposed to address this issue, the preservation of semantic information in speech has received relatively little attention. To achieve a balance between speaker de-identification and speech recognition performance, this study proposes an adversarial noise-based approach. The proposed approach is designed based on the differences in input processing between speaker and speech recognition systems. Experimental results show that the proposed method can achieve a meaningful trade-off between speaker de-identification performance and speech recognition performance, with a balance between the two observed particularly in the 10%–20% noise intensity range. These findings suggest that the proposed method can serve as a practical alternative for privacy-preserving speech security in cloud-based speech service environments.

▸ **Key words**: Speaker De-identification, Adversarial Attack, Privacy Preservation, Speaker Recognition, Speech Recognition

[요 약]

클라우드 기반 음성 인식 서비스는 높은 편의성과 접근성을 기반으로 다양한 분야에서 활용된다. 그러나 음성 데이터가 원격 서버에서 처리되는 과정에서 화자 정보 노출로 인한 개인 정보 보호 침해에 대한 우려도 제기되고 있다. 이러한 문제를 해결하기 위해 다양한 화자 비식별화 기법이 제안되었으나 음성 정보 보존에 대한 고려는 상대적으로 부족하였다. 이에 본 연구에서는 화자 비식별화와 음성 인식 성능 간의 균형을 달성하기 위해 적대적 노이즈 기반 화자 비식별화 기법을 제안한다. 해당 기법은 화자 인식과 음성 인식 간의 입력 처리 방식에 기반하여 설계되었다. 실험 결과, 제안 기법은 화자 비식별화와 음성 인식 성능 사이의 유의미한 트레이드오프 관계를 확인하였으며, 특히 10%-20% 노이즈 강도 구간에서 두 성능 간의 균형을 관찰하였다. 이는 제안된 기법이 클라우드 음성 서비스 환경에서 개인 정보 보호를 위한 실질적인 대응 수단으로 활용될 수 있음을 시사한다.

▸ **주제어**: 화자 비식별화, 적대적 공격, 개인 정보 보호, 화자 인식, 음성 인식

• First Author: Haram Kang, Corresponding Author: Kyungtae Kang
*Haram Kang (hrkang@hanyang.ac.kr), Dept. of Applied Artificial Intelligence, Hanyang University
**Sangwoon Yun (swyun@hanyang.ac.kr), Dept. of Computer Science and Engineering, Hanyang University
***Jemin Ahn (ahnjemin@hanyang.ac.kr), Research Institute of AI Convergence, Hanyang University
****Kyungtae Kang (ktkang@hanyang.ac.kr), Dept. of Artificial Intelligence, Hanyang University
• Received: 2025. 12. 01, Revised: 2026. 02. 20, Accepted: 2026. 03. 16.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

음성 인식은 인공지능 기술의 발전과 함께 다양한 산업 분야에서 핵심 기술로 자리잡고 있다. 시장 조사에 따르면, 글로벌 음성 인식 시장 규모는 2025년 약 87억 7천만 달러에서 2031년 약 236억 7천만 달러로 성장할 것으로 전망된다 [1]. 이러한 성장세는 스마트 스피커, 고객 센터 자동화 등 다양한 응용 분야에서 확인된다. 특히 최근에는 클라우드 환경을 기반으로 한 서비스 제공이 증가하면서 고성능 음성 인식 기능이 장소와 시간의 제약 없이 제공되어 사용자 편의성과 접근성이 크게 향상 되었다.

그러나 클라우드 기반 음성 인식 서비스의 확산은 새로운 보안 위협을 동반한다. 음성 데이터가 원격 서버로 전송되는 과정에서 화자 정보가 의도치 않게 노출될 수 있으며, 이는 개인정보 침해로 이어질 수 있다. 특히 음성 데이터는 화자의 신원과 발화 습관과 같은 민감한 정보가 포함되어 있어, 악용될 경우 심각한 보안 위협으로 이어질 수 있다. 실제로 최근 기업의 금융 관리자의 음성이 합성되어 약 1,850만 달러 규모의 암호화폐 사기에 악용된 사례가 보고되었다 [2]. 이러한 사례는 음성 데이터가 단순한 발화 정보가 아닌 화자를 특정할 수 있는 민감한 생체 정보로 기능함을 보여준다.

이러한 보안 위협에 대응하기 위한 방안으로 화자 비식별화(Speaker De-identification) 기술에 관한 연구가 활발히 이루어져 왔다. 대표적으로 스펙트럼 변형 [3], 임베딩 조작 [4], 음성 전환 기반 기법 [5] 등이 제안되었다. 이러한 접근법들은 신호를 재합성하거나 화자 특성을 변형하는 방식을 활용하여 화자 정보 제거에는 효과적이었으나 음성 신호의 구조적 변화로 인해 음성 인식 성능 저하가 발생하는 한계를 지녔다. 즉, 화자 정보 보호와 음성 인식 성능 간 균형은 여전히 해결되지 않은 핵심 과제로 남아 있으며, 실제 환경에서 활용되기 위해서는 화자 특성은 효과적으로 교란하면서도 발화 내용은 최대한 보존할 수 있는 새로운 방법이 요구된다.

이에 본 연구는 적대적 노이즈(Adversarial Noise) 기반의 화자 비식별화 기법을 제안한다. 제안된 기법은 가우시안 노이즈(Gaussian Noise) 및 라플라시안 노이즈(Laplacian Noise)를 활용하여 화자 정보를 교란하는 방식으로 설계되었으며, 화자 인식 정확도를 저하시킴과 동시에 음성 인식 성능 저하는 최소화하는 것을 목표로 한다. 이를 검증하기 위해 다양한 구조의 화자 인식 모델과 음성 인식 모델을 대상으로 실험을 수행하였으며, 제안 기법이 클라우드 기반 음성 서비스 환경에서 개인정보 보호 수단으로 활용 가능성을 분석하였다.

본 연구의 주요 기여는 다음과 같다.

- 본 연구는 화자 인식과 음성 인식 모델의 입력 처리 방식 차이에 기반하여 사용자 단에서 음성에 적대적 노이즈를 삽입하는 화자 비식별화 기법을 제안하였다.
- 본 연구는 다양한 화자 인식 및 음성 인식 모델과 상이한 노이즈 조건에서의 비교 실험을 통해, 제안 기법이 화자 비식별화 성능과 음성 인식 성능 유지 사이의 트레이드오프 관계를 분석하였다.
- 본 연구는 이러한 트레이드오프 관계를 바탕으로 제안 기법이 클라우드 음성 서비스 환경에서 실질적인 화자 비식별화 대안으로 활용될 수 있음을 시사하였다.

II. Preliminaries

1. Adversarial Noise

적대적 공격(Adversarial Attack)은 머신러닝 모델의 오분류를 유도하기 위해 입력 데이터에 미세한 변형을 가하는 기법을 의미한다 [6]. 일반적으로 적대적 공격은 모델의 기울기 정보를 활용하여 최적의 섭동을 생성하는 방식에 초점을 두어 왔다. 그러나 본 연구는 기울기 기반 최적화 방식 대신 모델이 구조적으로 미세한 입력 변형에 민감하게 반응할 수 있다는 점에 주목하여 확률적 교란 방식을 적용하였다. 이때, 해당 교란을 적대적 노이즈로 정의하였으며, 통계적 확률 분포를 따르는 가우시안 노이즈와 라플라시안 노이즈를 적대적 노이즈로 활용하였다.

가우시안 노이즈는 평균을 중심으로 대칭적인 정규 분포를 따를 통계적 노이즈로, 신호 처리 분야에서 신호의 세부 특성에 미세한 변화를 유도하면서도 전체적인 구조를 유지하는 특성을 가진다. 또한, 해당 노이즈는 자연계의 무작위 잡음을 모델링하는데 적합하며, 통신 시스템의 가산 백색 가우시안 노이즈(AWGN, Additive White Gaussian Noise) 분석 모델로 사용된다 [7]. 가우시안 노이즈의 확률 밀도 함수는 수식 (1)과 같이 정의된다.

$$GN(x|\mu,\sigma) = \frac{1}{(\sqrt{2\pi\sigma^2})} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (1)$$

가우시안 노이즈의 강도는 분산 σ^2를 조절하여 제어되며, 분산이 커질수록 교란의 범위와 세기가 증가한다.

라플라시안 노이즈는 이중 지수 분포를 따르는 통계적 노이즈로, 신호 처리 분야에서 불연속적이며 예측이 어려운 교란을 유도하며 가우시안 노이즈보다 더 뚜렷한 왜곡

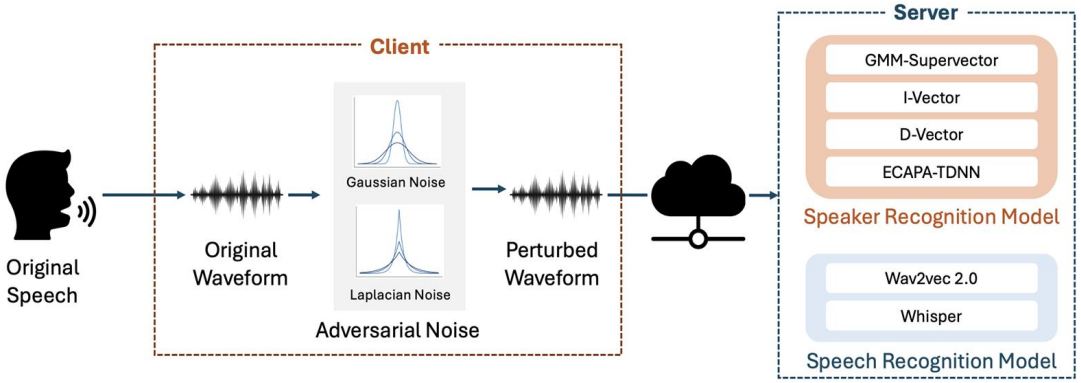

Fig. 1. Adversarial Noise-based Speaker De-identification

을 발생시킨다. 이러한 특성으로 인해 차등 프라이버시(Differential Privacy)에서는 개별 데이터에 대한 정보 노출을 완화하기 위한 대표적 기법인 라플라스 메커니즘에 라플라시안 노이즈가 사용된다 [8]. 라플라시안 노이즈의 확률 밀도 함수는 수식 (2)과 같이 정의된다.

$$LN(x|\mu,b) = \frac{1}{2b}\exp^{(-\frac{|x-\mu|}{b})} \qquad (2)$$

이때, 라플라시안 노이즈의 강도는 스케일 파라미터 b에 의해 조절되며, 스케일 파라미터가 증가할수록 교란 강도가 높아진다.

2. Speaker De-identification

화자 비식별화는 화자를 특정할 수 있는 음향적 특성을 변형하는 기술을 의미한다. 음성 데이터에 포함된 화자 정보 보호의 필요성이 제기되면서 다양한 화자 비식별화 기법이 제안되어 왔다.

Wang et al. [9]은 종단형 화자 인식 모델에 적대적 교란을 적용하여 화자 인식 성능이 인위적으로 저하될 수 있음을 입증하였다. Costandache et al. [3]은 피치 조절을 통해 화자 고유 특성을 약화시키는 신호 처리 기반 비식별화 방법을 제안하였으며, 실험을 통해 그 효과를 검증하였다. 그러나 음성 인식 성능은 고음역 변환 시 단어 오류율(WER, Word Error Rate)이 최대 0.33, 저음역 변환 시 최대 0.47로 증가하며 성능 저하가 발생하였다. Tavi et al. [4]은 포먼트 변조와 기본 주파수 궤적 조작을 결합한 비식별화 방법을 제시하였다. 음성 명료도 지표인 STOI (Short-Time Objective Intelligibility)를 활용하여 음성 인식 품질을 평가한 결과, 0.64-0.71 수준의 수치를 보이며 음성 인식 성능 저하가 관찰되었다. 또한, Yuan et al. [5]은 변분 인코더를 활용하여 실제 화자를 가상의 화자로 변환하는 zero-shot 기반 화자 비식별화 기법을 제안하였으며, 이때 WER은 약 0.30-0.35 수준으로 보고 되었다.

또한, Woszczyk et al . [10]은 치매 진단을 위한 음성 데이터 보호를 위해 도메인 지식 기반의 특징 분리 기법을 활용하여 신원 정보를 제거하고자 하였다. 실험 결과, 화자 인식 F1-score를 0.01%까지 낮추며 강력한 화자 비식별화 성능을 달성하였으나 특징 분리 과정에서 발생한 신호 왜곡으로 인해 WER은 최대 0.35까지 상승하는 것을 확인하였다. Tomashenko et al. [11]은 다중 화자 대화 환경에서 특정 타겟 화자 익명화를 위해 TSA(Target Speaker Anonymization) 프레임워크를 제시하였다. 해당 연구는 화자 추출과 변환 과정을 결합하여 비식별화 시나리오를 구현하였으나 신호 분리 및 재합성 과정에서 발생하는 아티팩트로 인해 음성 인식 성능이 저하되는 결과를 보였다. 실험 결과, 익명화된 음성을 일반 음성 인식 모델로 처리할 경우, WER이 최대 0.41까지 증가하였다.

이처럼 기존 연구들은 다양한 방식으로 화자 비식별화를 실현하였으나, 음성 인식 성능 유지 측면은 상대적으로 제한적으로 다루어졌다. 특히 실제 응용 환경에서 요구되는 인식 성능을 안정적으로 확보하기 어렵다는 것을 확인하였다. 이에 본 연구는 화자 정보를 효과적으로 비식별화하는 동시에 음성 인식 성능을 실용적인 수준으로 유지할 수 있는 방안을 제안한다.

III. Proposed Method

본 연구에서는 클라우드 기반 음성 인식 서비스 환경에서 발생할 수 있는 화자 정보 노출을 완화하는 동시에 음성 인식 품질을 유지하기 위한 방안으로 적대적 노이즈 기반 화자 비식별화 기법을 제안한다.

본 연구의 위협 시나리오는 사용자의 음성이 원격 서버로 전송되거나 처리되는 과정에서 외부에 노출되고, 공격자가 이를 탈취하여 화자 정보를 추출하거나 화자 인증 기

반 서비스 우회에 악용할 수 있는 상황으로 설정한다. 특히 탈취된 음성이 화자 인증 기반 보안 서비스에 활용되더라도 해당 음성이 원화자로 확인되지 않아야 한다.

이를 위해 본 연구는 별도의 재합성 없이 음성에 적대적 노이즈를 삽입하는 방식을 채택하였다. 이 방식은 복잡한 생성 과정 없이 사용자 단에서 전처리 형태로 적용될 수 있으며, 서버로 음성이 전송되기 이전 단계에서 화자 정보 보호를 수행한다는 특징을 갖는다.

이러한 접근은 화자 인식 모델과 음성 인식 모델의 입력 처리 방식 차이에 근거한다. 일반적으로 화자 인식 모델은 MFCC(Mel Frequency Cepstral Coefficient)와 같은 음향 특징 벡터를 활용하여 화자의 고유한 음성 특성을 추출하므로 미세한 교란에도 성능 저하가 발생하기 쉽다 [9]. 반면, 최신 음성 인식 모델은 원시 파형을 입력으로 받아 엔드투엔드 방식으로 특징을 추출하고 문맥 기반 구조를 통해 해석하기 때문에 일정 수준의 교란에도 비교적 강건한 성능을 유지할 수 있다 [12].

제안된 기법의 유효성을 검증하기 위해, 본 연구는 Fig. 1과 같이 노이즈가 삽입된 음성이 원격 서버로 전송되어 처리되는 클라우드 환경 기반의 흐름을 구성하였다. 구체적으로, 원본 음성에 가우시안 노이즈 및 라플라시안 노이즈를 삽입하여 화자 정보가 교란된 음성을 생성하고, 이를 음성 인식 및 화자 인식 관점에서 각각 분석하였다. 음성 인식 측면에서는 노이즈가 삽입된 음성이 실제 서비스 기능으로 활용 가능한지와 제안 기법 적용 이후에도 전사 성능이 실용 가능한 수준으로 유지되는지를 확인하였다. 반면, 화자 인식 측면에서는 원격 서버로 전송 및 처리되는 과정에서 음성이 악용될 수 있는 상황을 고려하여, 노이즈가 삽입된 음성이 공격자에 의해 사용되더라도 원화자로 식별되지 않는지를 검토함으로써 제안 기법의 화자 정보 보호 효과를 평가하였다. 구체적인 실험 절차는 이후 절에서 상세히 기술한다.

1. Description of Dataset

실험에는 LibriSpeech train-clean-360 [13] 데이터셋을 사용하였다. 해당 데이터셋은 약 360시간 분량의 영어 오디오북 음성으로 구성되어 있으며, 총 921명의 화자에 대한 음성 및 전사 정보를 포함한다.

전체 104,014개의 음성 파일에 대해 화자 단위 분할을 적용하여, 각 화자별 발화 개수의 10%를 평가용으로 분리하였다. 이를 통해 모든 화자가 학습 데이터와 평가 데이터에 균형 있게 포함되도록 하였다. 최종적으로 93,612개의 학습 데이터와 10,402개의 평가 데이터를 구성하였다.

2. Preprocessing of Dataset

화자 인식 실험을 위해 음성 데이터로부터 13차원의 MFCC를 추출하였다. 추출 과정에서는 16kHz로 샘플링된 음성을 입력으로 하였으며, librosa 라이브러리를 활용하여 각 프레임 단위의 MFCC를 계산한 뒤 정규화를 수행하였다. 이때, 발화별 길이 차이를 보정하기 위해 Zero-padding을 적용하여 모든 입력 특징 벡터가 동일한 길이를 갖도록 전처리 작업을 수행하였다.

3. Generation of Adversarial Noise

본 연구에서는 적대적 노이즈가 화자 인식 모델과 음성 인식 모델에 미치는 반응 차이를 분석하기 위해 원본 음성 신호에 가우시안 노이즈 및 라플라시안 노이즈를 각각 삽입하여 교란 음성을 생성하였다.

이때 서로 다른 통계적 특성을 갖는 노이즈 유형과 다양한 노이즈 강도를 설정함으로써, 노이즈 분포 특성과 강도 변화에 따른 화자 비식별화 효과 및 음성 인식 성능 변화를 단계적으로 분석하고자 하였다.

두 노이즈의 평균은 0으로 고정하였으며, 강도는 원본 음성 신호의 표준편차를 기준으로 한 상대적인 비율로 조절하였다. 즉, 노이즈의 크기는 각 음성 신호의 변동성에 비례하도록 설정되어 서로 다른 크기를 갖는 음성에 대해서도 일관된 수준의 교란이 적용되도록 하였다. 본 연구에서 사용한 노이즈의 강도는 0%부터 100%까지 10% 단위로 증가시키는 방식으로 설정하였다. 예를 들어, 노이즈 강도 10%는 원본 음성 신호 표준편차의 약 10%에 해당하는 크기의 노이즈가 적용되었음을 의미하며, 50%는 표준편차의 약 절반에 해당하는 노이즈가 삽입되었음을 의미한다. 이렇게 생성된 노이즈는 원본 음성 파형에 가산하는 방식으로 적용되었다.

4. Training of Speaker Recognition Model

본 단계에서는 적대적 노이즈의 화자 비식별화 효과가 특정 구조의 화자 인식 모델에 한정되지 않고 나타나는지를 확인하기 위해 다양한 화자 인식 모델을 학습하였다. 이에 따라 통계 기반 모델과 DNN(Deep Neural Network) 기반 모델을 사용하여 화자 인식 성능에 미치는 영향을 비교 및 분석하였다.

GMM-Supervector [14]는 GMM(Gaussian Mixture Model) 기반의 전통적인 화자 인식 기법이다. 해당 기법은 UBM(Universal Background Model)을 MAP 적응하여 각 화자별 평균 벡터를 추출하고 해당 벡터를 연결하여 고차원 벡터를 생성한다. 본 연구에서는 해당 모델을 가장

단순한 통계 기반 접근법으로 설정하여, 노이즈가 화자 인식에 미치는 영향을 통계적 수준에서 분석하기 위한 모델로 활용하였다. 이를 위해 128개의 혼합 성분을 갖는 UBM을 활용하여 벡터를 생성하였다.

I-Vector [15]는 TV(Total Variability) 공간을 기반으로 고정 길이의 화자 임베딩을 생성한다. TV는 화자 변동성과 채널 변동성을 포함하는 저차원 공간으로, 두 요인을 설명할 수 있는 통계적 표현을 제공한다. 본 연구에서는 I-Vector를 통계 기반 접근의 확장형 모델로 채택하여 GMM-Supervector보다 더 압축된 표현이 노이즈에 대해 어느 정도의 강건성을 보이는지를 평가하였다. 실험에서는 128개의 혼합 성분을 갖는 UBM을 기반으로 TV 행렬을 학습한 뒤 400차원의 I-Vector를 추출하였다.

D-Vector [16]는 DNN 기반 초기 화자 인식 모델로 프레임 단위의 임베딩을 추출한 뒤 평균 풀링을 통해 발화 수준의 화자 벡터를 구성한다. 이는 기존의 통계 기반 방식과 달리 DNN을 이용한 학습 표현을 활용된다는 점에서 차별화된다. D-Vector는 이러한 차이에 기반하여 통계적 모델과 DNN 모델 간의 성능 차이 및 노이즈 반응 특성을 비교하기 위한 모델로 사용하였다. 학습 과정에서는 CrossEntropyLoss를 손실 함수로 사용하고 Adam 옵티마이저와 Softmax 분류기를 적용하였다.

ECAPA-TDNN [17]은 TDNN(Time-Delay Neural Network)을 기반으로 다양한 시간 스케일의 정보를 통합하고 채널 어텐션 메커니즘을 적용하여 화자 특성을 정교하게 추출하는 DNN 기반 화자 인식 모델이다. 또한, Self-attentive Pooling을 통해 발화 수준의 임베딩을 집약하여 표현력을 강화한다. 본 연구에서는 해당 모델을 가장 성능이 높은 화자 인식 모델로 설정하여 정교한 구조를 가진 최신 신경망에서도 제안한 노이즈가 화자 인식 성능에 유의미한 영향을 미치는지를 검증하고자 하였다. 학습 과정에서는 손실 함수로 CrossEntropyLoss를 사용하고 Adam 옵티마이저와 Softmax 분류기를 적용하였다.

5. Training of Speech Recognition Model

본 단계에서는 적대적 노이즈가 음성 인식 성능에 미치는 영향을 분석하기 위해 대표적인 최신 음성 인식 모델인 Wav2vec 2.0 [12]과 Whisper [18]를 대상으로 실험을 진행하였다. 서로 다른 특성을 가진 두 모델을 사용하여 제안 기법의 영향을 다양한 음성 인식 구조에서 비교 및 분석하고자 하였다.

Wav2vec 2.0은 자기 지도 학습을 기반으로 사전 학습된 음성 인식 모델이다. 해당 모델은 CNN 기반의 특징 추출기를 통해 음성 특징을 추출하고 트랜스포머 기반의 컨텍스트 네트워크를 통해 전반적인 발화의 맥락을 이해한다. 음성 인식 분야에서 활용되는 대표적인 모델로, 노이즈 삽입 전후의 인식 성능 변화를 정량적으로 평가하기 위해 사용하였다. 구현은 Fairseq 프레임워크를 기반으로 진행하였으며 학습 데이터셋을 활용하여 추가적인 미세 조정을 수행하였다.

Whisper는 약지도 학습 기반의 음성 인식 모델로 인코더-디코더 구조의 Transformer를 기반으로 다양한 언어와 잡음 환경을 포함한 대규모 코퍼스를 학습하였다. 본 연구에서는 Whisper를 대규모 사전 학습을 통해 일반화된 음성 표현을 학습한 강건한 모델로 설정하여, 적대적 노이즈 삽입 이후의 음성 인식 성능을 평가하기 위한 모델로 활용하였다. 구현은 HuggingFace의 Transformers 라이브러리를 기반으로 수행하였다.

IV. Experiment and Results

본 실험은 제안된 적대적 노이즈 기반 화자 비식별화 기법의 성능을 검증하기 위해 수행되었다. 이를 위해 가우시안 노이즈(GN) 및 라플라시안 노이즈(LN)를 적용하여, 노이즈 삽입이 화자 인식 및 음성 인식 모델의 성능에 미치는 영향을 정량적으로 분석한다.

또한, 다양한 강도의 노이즈를 구성하여 강도 변화에 따른 성능 변화를 비교함으로써 화자 비식별화 효과와 음성 인식 성능 간의 트레이드오프 양상을 분석하고자 하였다.

다만, 본 분석은 본 연구에서 설정한 데이터 및 실험 조건을 기준으로 수행되었으며, 해당 범위 내에서의 성능 변화를 중심으로 결과를 해석하였다.

1. Experimental Setup

본 절에서는 제안 기법의 성능 평가를 위한 실험 환경 및 음성 인식 모델의 미세 조정 설정을 기술한다.

하드웨어 환경으로는 NVIDIA GeForce RTX 3080 GPU를 사용하였으며 실험에 활용된 주요 소프트웨어 프레임워크 및 라이브러리 상세 명세는 Table 1과 같다.

모델 학습 및 평가에는 LibriSpeech Clean 데이터셋을 활용하였다. 음성 인식 모델은 사전 학습된 Wav2vec 2.0 및 Whisper을 미세 조정하였다. 학습 시 적용된 하이퍼파라미터의 설정값은 Table 2에 정리하였다.

Table 1. Implementation Details

Category	Library	Version
Language	Python	3.9
Framework	Pytorch	2.5.1
Numerical Computation	Numpy	2.2.0
Audio Processing	Librosa	0.11.0
	Soundfile	0.13.1

Table 2. Fine-tuning Configuration of Speech Recognition Models

	Wav2vec 2.0	Whisper
Pretrained Model	wav2vec2-base	whisper-small
Optimizer	AdamW	AdamW
LR Scheduler	Linear Decay	Linear Decay
Batch Size	16	16
Learning Rate	1e-4	1e-5
Training Steps / Epochs	30 epochs	40,000 steps
Precision	fp16	fp16

2. Speaker Recognition

본 절에서는 제안된 기법이 화자 식별 가능성에 미치는 영향을 확인하고 모델 구조에 따라 나타나는 노이즈 민감도 차이를 분석한다. 이를 위해 각 모델에서 추출한 화자 임베딩과 화자별 대표 벡터 간의 코사인 유사도를 계산하고, 가장 높은 유사도를 보이는 화자를 최종 예측 결과로 결정하였다. 이후 화자 예측 정확도를 주요 지표로 활용하여 노이즈 삽입이 비식별화 성능에 미치는 영향을 정량적으로 평가하였다.

Table 3에서 확인되는 바와 같이 모든 모델에서 노이즈 강도가 증가할수록 화자 인식 정확도가 감소하는 경향이 관찰되었으며, 그 감소 폭은 모델의 구조적 특성에 따라 다르게 나타났다.

통계 기반 화자 인식 모델에서는 특징 표현 방식과 정보 압축 정도에 따라 노이즈에 대한 반응 양상이 다르게 나타났다. 20% 노이즈 강도 조건에서 GMM-Supervector는 평균 15.6%의 정확도를 보여 동일한 통계 기반 모델인 I-Vector보다 상대적으로 높은 성능을 유지하였다. 이는 고차원 공간에서 통계적 특징을 반영하는 표현 방식이 노이즈가 삽입된 조건에서도 일정 수준의 정보를 유지했기 때문으로 해석된다. 반면, 압축된 표현의 강건성을 평가하기 위해 사용한 I-Vector는 동일 조건에서 평균 정확도가 약 3.2%까지 감소하여 화자 인식 모델 가운데 노이즈의 영향에 가장 민감한 경향을 보였다. 이러한 결과는 저차원 압축 과정에서 발생하는 정보 손실이 입력 교란에 대한 민감도와 관련될 가능성을 시사한다. 종합하면 제안한 기법의 화자 비식별화 효과는 통계 기반 모델 내에서도 동일하게 나타나기보다 각 모델의 특징 표현 구조에 따라 다르게 형성될 수 있음을 보여준다.

Table 3. Effect of Noise Level on Speaker Recognition Accuracy (%)

Model	Noise	0	10	20	30	40	50
GMM Supervector	GN	97.0	35.3	18.4	11.9	8.3	6.0
	LN	96.0	25.7	12.8	7.8	5.1	3.5
I-Vector	GN	84.4	8.3	3.8	2.5	2.0	1.7
	LN	84.4	5.5	2.7	2.0	1.4	1.2
D-Vector	GN	95.0	33.4	17.8	12.0	9.2	7.1
	LN	91.6	24.9	13.0	8.4	6.0	4.9
ECAPA -TDNN	GN	95.2	54.6	32.8	22.9	17.1	13.4
	LN	90.2	43.8	24.3	16.0	11.7	8.7
Average		91.7	28.9	15.7	10.4	7.6	5.8

DNN 기반 모델에서도 구조적 차이에 따른 반응 차이가 관찰되었다. D-Vector는 프레임 별 임베딩을 단순 평균하여 화자 표현을 구성하는 특성으로 인해 노이즈 강도 증가에 따라 성능 저하가 비교적 크게 나타났다. 특히 20% 노이즈 조건에서 평균 정확도가 15.4%까지 감소한 반면 ECAPA-TDNN은 동일한 조건에서 평균 28.5%의 정확도를 기록하여 화자 인식 실험 대상 중 상대적으로 높은 성능을 유지하였다. 이는 채널 어텐션 기반의 특징 선택 메커니즘이 노이즈 환경에서도 화자 식별에 유효한 핵심 특징 패턴에 집중함으로써 정보를 보다 안정적으로 반영한 결과로 볼 수 있다. 이러한 결과는 DNN 기반 모델 내에서도 임베딩 집계 방식과 특징 선택 구조에 따라 노이즈 민감도가 다르게 나타날 수 있음을 보여준다.

종합적으로 화자 인식 모델의 성능 변화는 통계 기반이나 DNN 기반이라는 범주적 구분보다는 정보의 압축 방식이나 내부 구조적 특성과 보다 밀접하게 관련되어 나타나는 경향을 보였다. 또한, 이러한 구조적 차이에도 불구하고 모든 실험 대상에서 노이즈 강도 증가에 따른 화자 예측 정확도 저하가 공통적으로 관찰되었다. 이는 제안 방식이 서로 다른 구조의 화자 인식 모델에서도 일관된 성능 저하를 유도할 수 있음을 보여주며 다양한 환경에서 화자 비식별화에 활용될 가능성을 시사한다.

다만, 본 절의 결과는 Top-1 정확도 기반의 화자 식별(Identification) 관점에서 성능 변화를 중심으로 확인한 것이며, 등오류율(EER, Equal Error Rate)와 같은 화자 검증(Verification) 관점에서의 추가적인 검토가 필요하다.

Table 4. Effect of Noise Level on Speech Recognition Word Error Rate (WER)

Model	Noise	0	10	20	30	40	50
Wav2vec 2.0	GN	0.15	0.18	0.24	0.33	0.44	0.56
	LN	0.15	0.20	0.30	0.45	0.61	0.75
Whisper	GN	0.03	0.04	0.05	0.06	0.07	0.09
	LN	0.03	0.05	0.06	0.08	0.09	0.10
Average		0.09	0.12	0.17	0.23	0.32	0.40

3. Speech Recognition

본 절에서는 제안된 기법이 음성 인식 성능에 미치는 영향을 확인하고 모델 별 강건성 차이를 분석한다. 이를 위해 노이즈가 삽입된 음성을 음성 인식 모델에 입력하여 전사 결과를 생성하였으며 이를 정답 텍스트와 비교하여 WER을 산출하였다. WER은 수치가 낮을수록 원본 음성의 언어적 정보가 보존됨을 의미하며 본 실험에서는 이를 통해 음성 인식 성능을 정량적으로 분석하였다.

Table 4에 제시된 바와 같이, 노이즈 강도가 증가함에 따라 모든 모델의 WER이 점진적으로 상승하는 경향을 보였다. 그러나 화자 인식의 성능 저하 양상과 비교할 때 음성 인식 모델은 상대적으로 완만한 성능 변화를 보였다. 특히 20% 강도의 노이즈가 삽입된 조건에서도 전체 모델의 평균 WER은 0.17 수준으로 나타났다. 이는 본 연구의 실험 환경에서는 일정 수준의 노이즈가 삽입되더라도 실제 응용 서비스에서 허용 가능한 수치인 WER 0.2 이하를 안정적으로 유지될 수 있음을 보여준다 [19].

이러한 결과는 음성 인식 모델의 구조적 특성과 학습 방식과 관련이 있는 것으로 해석된다. 본 연구에서 활용된 Wav2vec 2.0과 Whisper는 모두 트랜스포머 기반 구조를 사용하며 발화 전반의 문맥 정보를 반영할 수 있는 특성을 가진다. 이에 따라 음향 신호의 일부가 노이즈에 의해 왜곡되더라도 문맥적 단서를 바탕으로 언어 정보를 보완한 것으로 보인다. 이는 음향적 특징에 민감한 화자 인식 모델과 비교할 때 음성 인식 모델이 교란에 대해 상대적으로 완만한 성능 저하를 보인 결과와도 연결된다.

또한, 모델 간 비교에서는 Whisper가 Wav2vec 2.0보다 동일한 노이즈 환경에서 더 낮은 WER을 유지하는 경향을 보였다. 20% 노이즈 조건에서 Whisper의 평균 WER은 약 0.05인 반면에 Wav2vec 2.0은 평균 0.27의 WER을 나타냈다. 이러한 차이는 사전 학습 데이터의 규모와 다양성 및 잡음에 대한 노출 정도와 관련이 있는 것으로 분석된다. 특히 Whisper는 다양한 조건을 포함하는 대규모 데이터로 학습되었기 때문에 교란이 존재하는 환경에서도 언어 정보 보존 측면에서 상대적으로 안정적인 성능을 보인 것으로 해석된다.

종합적으로, 음성 인식 실험 결과는 일정 수준의 노이즈가 추가된 조건에서도 언어적 정보가 비교적 유지될 수 있음을 보여준다. 이는 제안된 기법이 화자 비식별화 효과를 보이는 동시에 음성 인식 성능은 상대적으로 완만하게 저하될 수 있음을 의미한다.

다만, 본 절의 결과는 WER 기반의 객관적 성능 변화를 중심으로 확인한 것이며 실제 청취 품질이나 인지적 발화 이질성에 대한 평가는 향후 주관적 평가를 통해 추가적으로 검토할 필요가 있다.

4. Comprehensive Analysis

본 절에서는 앞선 실험 결과를 바탕으로 서로 다른 통계적 특성을 갖는 노이즈 조건에서 나타나는 경향을 비교하고 화자 비식별화 효과와 음성 인식 성능 간의 트레이드오프에 대해 종합적으로 분석한다.

4.1 Comparison of Gaussian and Laplacian Noise

본 연구에서는 서로 다른 통계적 특성을 갖는 가우시안 노이즈와 라플라시안 노이즈를 적용하여 제안 기법이 노이즈 유형에 따른 성능 차이를 비교하였다.

실험 결과, 두 노이즈 유형 모두에서 화자 인식 성능 저하와 음성 인식 성능 변화가 함께 관찰되었으나, 세부적인 변화 양상에서는 각 노이즈의 분포 특성에 따라 차이가 나타났다. 라플라시안 노이즈는 가우시안 노이즈에 비해 상대적으로 더 불연속적이고 큰 변동을 포함하는 교란을 유도한다. 이러한 특성이 실제 실험에서도 모델의 화자 인식 정확도와 음성 인식 성능에 반영된 것을 확인하였다. 특히 라플라시안 노이즈 환경에서는 전반적으로 화자 인식 정확도와 음성 인식 성능이 가우시안 노이즈 대비 다소 낮게 나타나는 경향이 관찰되었다. 이러한 결과는 노이즈의 통계적 특성이 모델의 반응 양상에 영향을 미칠 수 있음을 시사한다.

종합하면, 제안 기법은 특정 노이즈 유형에만 한정된 결과를 보이기보다는 서로 다른 통계적 특성을 갖는 노이즈 조건에서도 공통적인 성능 변화 경향을 나타냈다. 동시에 노이즈의 통계적 특성에 따라 세부적인 성능 변화 양상에는 차이가 나타날 수 있음을 확인하였다. 이는 적대적 노이즈 기반 방식이 노이즈 유형이 달라지더라도 화자 비식별화 효과를 유도할 수 있음을 보여준다. 또한, 이러한 특성 차이를 반영한 다양한 노이즈 설계 및 적용은 향후 성능 변화를 보다 정교하게 조절하는 방향으로 확장될 수 있음을 시사한다.

4.2 Exploration of a Practical Trade-off Range

본 연구는 제안 기법이 사용자의 화자 정보를 비식별화하면서 음성 인식 서비스의 유용성을 유지할 수 있는지를 살펴보기 위해 노이즈 강도에 따른 성능 변화를 비교하였다. 이때, 음성 인식 성능의 유용성을 판단하기 위한 참고 기준으로 평균 WER 0.2 이하 조건을 함께 고려하였다.

분석 결과, 노이즈 강도 10%-20% 구간에서는 화자 비식별화 성능과 음성 인식 성능 간의 트레이드오프 관계가 관찰되었다. 구체적으로 10% 노이즈 조건에서는 음성 인식 모델의 평균 WER이 0.12로 비교적 낮은 수준을 유지하였으나 ECAPA-TDNN은 49.2%의 화자 인식 정확도를 보여 화자 비식별화 효과는 제한적으로 나타났다. 반면, 20% 노이즈 조건에서는 화자 인식 모델의 평균 정확도가 15.7%까지 감소하여 화자 비식별화 효과가 보다 뚜렷하게 나타났고 음성 인식 성능도 전체 평균 WER 0.17 수준으로 유지되었다. 다만, Wav2vec 2.0에서는 WER이 0.30까지 증가한 경우도 확인되어 해당 구간은 모델에 따라 성능 차이가 존재하는 범위로 해석할 필요가 있다.

이러한 결과를 종합하면, 노이즈 강도 10%-20% 구간은 본 연구의 실험 환경에서 화자 비식별화 효과와 음성 인식 성능 보존 간의 트레이드오프가 관찰된 범위로 해석할 수 있다. 이는 제안 기법을 통해 두 성능 간의 균형이 형성될 수 있음을 의미하며 제안 기법이 화자 정보 보호와 음성 인식 기반 서비스의 유용성 유지를 함께 고려할 수 있는 방안으로 활용될 수 있음을 시사한다.

V. Discussion

본 연구에서는 적대적 노이즈를 활용하여 음성 인식 성능을 안정적으로 유지하면서 화자 비식별화를 달성할 수 있는 기법을 제안하고 실험을 통해 그 유효성을 검증하였다. 본 장에서는 이러한 실험 결과를 바탕으로 본 연구의 한계와 의의를 논의하고 향후 연구 방향을 제시한다.

1. Dataset Constraints and Generalization

본 연구는 음성 연구 분야의 대표적인 벤치마크인 LibriSpeech 데이터셋을 활용하여 제안한 적대적 노이즈 기반 화자 비식별화 기법의 유효성을 검증하였다. 이는 통제된 환경 내에서 제안된 기법의 기초 성능과 기술적 타당성을 확보했다는 점에서 의의를 갖는다.

그러나 해당 데이터셋은 소음이 배제된 영어 낭독 기반의 음성으로 구성되어 있어, 실제 운용 환경의 다양한 변동 요인을 충분히 반영하기에는 한계가 있다. 특히 실제 환경에서는 자발적 발화, 배경 소음, 채널 왜곡 등이 복합적으로 작용한다. 따라서 향후 연구에서는 다국어 및 다도메인 데이터를 포함한 다양한 발화 스타일과 비정형적인 소음 환경에서도 비식별화 성능이 유지되는지 확인하는 일반화 검증을 수행하고자 한다.

2. System Deployment and Pre-processing

본 연구에서 제안한 기법은 음성 데이터가 클라우드 서버로 전송되기 전 사용자 단에서 전처리가 수행되는 것을 전제로 한다. 구체적인 활용 사례로 IoT 기기의 임베디드 모듈에 본 기법을 통합할 수 있다. 이 경우 사용자가 명령어를 발화하는 즉시 기기 내부에서 적대적 노이즈를 주입하여 비식별화된 데이터를 생성한 뒤 클라우드 서버로 전송하는 방식으로 작동한다. 이러한 구조는 서비스 제공자의 내부 모델에 대한 접근 없이도 개인정보 노출 위험을 완화할 수 있어 클라이언트 측 제어가 가능한 모바일 및 IoT 환경에 실용적인 적용 대안이 될 수 있다.

그러나 실제 클라우드 음성 서비스 환경에서는 노이즈 억제, 음질 보정 등 서버 측 전처리 과정이 수반될 수 있다[20]. 이로 인해 사용자 단에서 삽입된 적대적 노이즈가 부분적으로 제거되거나 왜곡될 가능성이 존재한다. 본 연구에서는 이러한 전처리 파이프라인 전체를 실험적으로 분석하지는 않았으나, 향후 연구에서는 서버 단 전처리 환경에서도 비식별화 성능을 유지할 수 있는 강건한 노이즈 생성 기법으로의 확장이 필요하다.

또한, 실제 배포 환경에서는 기술적 성능 외에도 서비스 제공자의 운영 정책 및 품질 관리 요구 사항을 고려해야 한다. 사용자 입력 음성의 의도적 변형은 비정상 입력 탐지 시스템 및 서비스 품질 보증 지표에 영향을 미칠 수 있으며 배포 단계에서 서비스 제공자와의 정책적 합의 또는 가이드라인 설정을 요구할 수 있다. 본 연구는 이러한 정책적 쟁점을 직접 다루기보다 기술적 가능성 검증에 초점을 두었으나, 향후에는 보다 현실적인 적용 시나리오를 통해 제안 기법의 실질적인 활용 범위를 확장하고자 한다.

3. Evaluation Scope and Extended Analysis

본 연구는 화자 인식 실험에서 Top-1 정확도를 검증 지표로 활용하여 결과를 분석하였다. 실험 결과, 다양한 화자 인식 모델 구조에서 일관된 비식별화 성능을 확인하였으나 화자 식별 시나리오에 국한된 분석이라는 한계가 있다. 추후에는 EER과 같은 검증 지표를 도입하여 실제 화자 검증 환경에서의 프라이버시 보호 효과를 다각도로

규명할 필요가 있다.

또한, 음성 인식 분석은 객관적 지표인 WER를 활용하였다. 향후 연구에서는 객관적 평가를 넘어 실제 사용자를 대상으로 한 청취 기반 주관적 평가를 수행할 계획이다. 이를 통해 제안 기법이 인지적 발화 이질성에 미치는 영향을 분석하여 기법의 실용적 완성도를 검증하고자 한다.

4. Future Research Directions

본 연구는 음성 탈취 및 화자 정보 노출을 중심으로 위협 시나리오를 설정하였다. 그러나 실제 환경에서는 공격자의 지식 수준에 따라 다양한 위협 시나리오가 존재할 수 있다. 따라서 향후 연구에서는 다양한 공격 상황을 체계적으로 정의하고 각 상황에서의 비식별화 효과를 추가적으로 분석할 필요가 있다.

또한, 본 연구에서는 통계적 노이즈를 이용하여 확률적 교란이 프라이버시 보존에 기여할 수 있음을 분석하였다. 향후 연구에서는 본 실험을 통해 확인된 비식별화 성능과 음성 품질 간의 트레이드오프 관계를 바탕으로 발화 조건과 주변 환경에 따라 노이즈 강도를 조절하는 적응형 비식별화 기법으로 확장하고자 한다.

VI. Conclusions

본 연구는 클라우드 기반 음성 인식 서비스 환경에서의 화자 정보 노출로 인한 개인 정보 보호 문제를 완화하기 위해 적대적 노이즈 기반 화자 비식별화 기법을 제안하였다. 제안 기법은 음성 인식 모델과 화자 인식 모델 간의 입력 처리 방식의 차이를 기반으로 사용자 단에서 음성에 적대적 노이즈를 삽입하는 방식을 채택하였다.

이를 검증하기 위해 본 연구에서는 가우시안 노이즈와 라플라시안 노이즈를 적용하여 다양한 화자 인식 모델과 음성 인식 모델을 대상으로 실험을 수행하였다. 실험 결과, 제안 기법은 화자 인식 성능을 저하시키는 동시에 음성 인식의 성능은 비교적 안정적으로 유지하는 경향을 보였다. 특히 노이즈 강도 10%-20% 구간이 화자 비식별화 효과와 음성 인식 성능 보존 사이의 트레이드오프를 확인할 수 있는 범위로 관찰되었다.

이는 제안 기법을 통해 화자 인식 성능 저하와 음성 인식 성능 보존 사이에서 균형이 형성될 수 있음을 보여주는 결과로 해석할 수 있다. 또한 서로 다른 구조의 화자 인식 모델과 음성 인식 모델 전반에서 유사한 경향이 관찰되었다는 점을 기반으로 제안 기법이 클라우드 음성 서비스 환경에서 실질적인 개인정보 보호 수단으로 활용될 수 있음을 시사한다.

향후 연구에서는 보다 다양한 데이터셋과 실제 서비스 환경을 반영한 검증과 화자 검증 기반 평가 및 청취 기반 주관적 평가를 확장하고 노이즈 강도를 동적으로 조절하는 적응형 비식별화 기법으로 발전하고자 한다.

ACKNOWLEDGEMENT

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2022-00155885, Artificial Intelligence Convergence Innovation Human Resources Development (Hanyang University ERICA)) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2024-00431388, the Global Research Support Program in the Digital Field program)

REFERENCES

[1] Statista, "Speech Recognition," Statista Website, *https://www.statista.com/outlook/tmo/artificial-intelligence/computer-vision/speech-recognition/worldwide*.

[2] Reality Defender, "Understanding the $603,000 Problem: The Real Cost of Voice Fraud in Banks," Reality Defender Website, *https://www.realitydefender.com/insights/the-603-000-problem-real-cost-of-voice-fraud-in-banks*.

[3] M. A. Costandache, A. Iftee, and D. Gifu, "A Speaker De-identification System Based on Sound Processing," Proceedings of Information Systems Development, Valencia, Spain, Sept. 2021.

[4] L. Tavi, T. Kinnunen, and R. G. Hautamäki, "Improving speaker de-identification with functional data analysis of f0 trajectories," Speech Communication, Vol. 140, pp. 1-10, June 2022. DOI: 10.1016/j.specom.2022.03.010

[5] R. Yuan, Y. Wu, J. Li, and J. Kim, "DeID-VC: Speaker De-identification via Zero-shot Pseudo Voice Conversion," Proceedings of Interspeech 2022, pp. 2593-2597, Incheon, Korea, Sept. 2022. DOI: 10.21437/INTERSPEECH.2022-11036

[6] J. C. Costa, T. Roxo, H. Proença, and P. R. M. Inácio, "How

Deep Learning Sees the World: A Survey on Adversarial Attacks & Defenses," IEEE Access, Vol. 12, pp. 61113-61136, 2024. DOI: 10.1109/ACCESS.2024.3395118

[7] A. Trabelsi, O. Mohamed, and Y. Audet, “Robust parametric modeling of speech in additive white Gaussian noise,” Journal of Signal and Information Processing, 6(2), pp. 99-108, 2015, DOI: 10.4236/jsip.2015.62010

[8] C. Dwork and A. Roth, "The Algorithmic Foundations of Differential Privacy," Foundations and Trends in Theoretical Computer Science, Vol. 9, No. 3−4, pp. 211−407, Aug. 2014. DOI: 10.1561/0400000042

[9] Q. Wang, P. Guo, and L. Xie, "Inaudible Adversarial Perturbations for Targeted Attack in Speaker Recognition," Proceedings of Interspeech 2020, pp. 4228-4232, Shanghai, China, Oct. 2020. DOI: 10.21437/Interspeech.2020-1955

[10] D. Woszczyk, A. Ranya and D. Soteris, “Prosody-Driven Privacy-Preserving Dementia Detection.”, ArXiv, 2024. abs/2407.03470

[11] N. Tomashenko, J. Yamagishi, X. Wang, Y. Liu and E. Vincent, “Target speaker anonymization in multi-speaker recordings.”, ArXiv, 2025, abs/2510.09307

[12] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: a framework for self-supervised learning of speech representations," Proceedings of the 34th International Conference on Neural Information Processing Systems, pp. 12449-12460, Vancouver, Canada, Dec. 2020.

[13] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," Proceedings of 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5206-5210, South Brisbane, Australia, Apr. 2015. DOI: 10.1109/ICASSP.2015.7178964

[14] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," IEEE Signal Processing Letters, Vol. 13, No. 5, pp. 308-311, May 2006. DOI: 10.1109/LSP.2006.870086

[15] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," IEEE Transactions on Audio, Speech, and Language Processing, Vol. 19, No. 4, pp. 788-798, May 2011. DOI: 10.1109/TASL.2010.2064307

[16] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4052-4056, Florence, Italy, May 2014. DOI: 10.1109/ICASSP.2014.6854363

[17] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," Proceedings of Interspeech 2020, pp. 3830-3834, Shanghai, China, Oct. 2020. DOI: 10.21437/Interspeech.2020-2650

[18] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," Proceedings of the 40th International Conference on Machine Learning, Vol. 202, pp. 28492-28518, Honolulu, USA, July 2023.

[19] Microsoft Azure, "Test accuracy of a custom speech model", Microsoft Learn, *https://learn.microsoft.com/en-us/azure/ai-services/speech-service/how-to-custom-speech-evaluate-data?pivots=ai-foundry-portal*

[20] AWS, "Understanding VoiceFocus for the Amazon Chime SDK PSTN audio service", Amazon Chime SDK, *https://docs.aws.amazon.com/chime-sdk/latest/dg/voice-focus.html*

Authors

Haram Kang received her B.S. degree in Computer Science Engineering from Inha Technical College, Incheon, Korea, in 2024. She is currently a master student degree in Applied Artificial Intelligence at Hanyang University. Her research interests include the practical application of artificial intelligence, focusing on generative AI, large language model, and AI-driven cybersecurity.

Sangwoon Yun received his B.S. degree in Computer Science and Engineering from Inha Technical College, Incheon, Korea, in 2021. In 2023, he received the master degree in Artificial Intelligence (M.A.I.) at Hanyang University, and he is currently a doctoral candidate (Ph.D cand.). His research interests lie primarily in systems, including operating systems, real-time embedded systems, and highly dependable computing. His recent research interest is in the interdisciplinary area of cyber-physical systems.

Jemin Ahn received his B.S. degree in Computer Science and Engineering from Hanyang University, Ansan, South Korea, in 2017, and his Ph.D. degree in Computer Science and Engineering from Hanyang University, Seoul, South Korea, in 2025. He is currently a Post-doctoral Research Associate at the Research Institute of AI Convergence, Hanyang University. His research interests lie at the intersection of cybersecurity and deep learning, with a focus on developing practical methods for detecting and preventing cyber threats. His recent work centers on leveraging natural language processing models to advance attack detection techniques.

Kyungtae Kang received his B.S. degree in Computer Science and Engineering in 1999, and his M.S. and Ph.D. degrees in Electrical Engineering and Computer Science in 2001 and 2007, respectively, from Seoul National University, Seoul, Korea. From 2008 to 2010, he was a postdoctoral research associate at the University of Illinois at Urbana-Champaign, IL, USA. In 2011, he joined the Department of Computer Science and Engineering at Hanyang University, Korea and is currently a tenured professor in the Department of Artificial Intelligence. His research interests primarily focus on systems, including operating systems, mobile systems, distributed systems, and real-time embedded systems. His recent work delves into the interdisciplinary field of cyber-physical systems.