JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 27, No. 2, pp. 503-509, Feb. 2026

Check for updates

딥페이크 음성 식별 성능 개선을 위한 전역 주파수/시간 연관관계 학습 강화 연구

장 혜 준[1] · 조 서 진[1] · 김 연 수[1] · 박 태 정[2*]
[1]덕성여자대학교 사이버보안학과 학부 과정
[2]덕성여자대학교 디지털소프트웨어학부 교수

Enhancing Global Frequency-Time Correlation Learning for Improved Deepfake Audio Detection

Hyejun Jang[1] · Seojin Cho[1] · Yeonsoo Kim[1] · Taejung Park[2*]

[1]Bachelor's Course, Department of Cybersecurity, Duksung Women's University, Seoul 01369, Korea
[2]Professor, Department of Department of Engineering and Converged Technology, Duksung Women's University, Seoul 01369, Korea

[요 약]

최근 음성 합성 및 변조 기술의 급속한 발전으로 인해 딥페이크 음성을 악용한 보이스 피싱, 사기, 신원 도용 등의 위협이 증가하고 있다. 이에 따라 기존 화자 검증 시스템을 보완할 수 있는 음성 딥페이크 탐지 기술의 중요성이 부각되고 있다. 본 연구에서는 기존 Multi-View Collaborative Learning(MVCLN)의 구조를 기초로, 스펙트로그램 인코더를 기존 CNN에서 Vision Transformer(ViT)로 변경한 음성 딥페이크 탐지 모델을 제안한다. 제안 모델은 waveform 기반 Transformer와 ViT 기반 spectrogram 인코더를 병렬로 구성하고, 두 표현 간의 상호 보완적 특성을 효과적으로 융합하기 위해 Waveform-Spectrogram Fusion Module(WSFM)을 적용하였다. 실험 결과, 제안한 모델은 seen 환경에서 AUC 0.9999, EER 0.38%를 기록하였으며, unseen 환경에서도 AUC 0.9653으로 기존 MVCLN 보다 성능이 향상되었다.

[Abstract]

The rapid advancement of voice synthesis and modulation technologies has increased the threat of exploitation of deepfake audio for voice phishing, fraud, and identity theft. Consequently, the importance of deepfake audio detection technology that can complement existing speaker verification systems has been highlighted. In this study, we propose a deepfake audio detection model based on the Multi-View Collaborative Learning Network (MVCLN) architecture, with Vision Transformer (ViT) replacing CNN as the spectrogram encoder. The proposed model employs a parallel configuration of a waveform-based Transformer and a ViT-based spectrogram encoder, and applies a Waveform-Spectrogram Fusion Module (WSFM) to effectively fuse the complementary characteristics between the two representations. Experimental results demonstrate that the proposed model achieves an AUC of 0.9999 and an EER of 0.38% in seen scenarios, and an AUC of 0.9653 in unseen scenarios, showing improved performance compared to the original MVCLN.

색인어 : 음성 딥페이크 감지, 멀티뷰 학습, 비전 트랜스포머, 음성 보안, 멜 스펙트로그램
Keyword : Speech Deepfake Detection, Multi-View Learning, Vision Transformer, Audio Security, Mel-Spectrogram

http://dx.doi.org/10.9728/dcs.2026.27.2.503

Received 29 December 2025; **Revised** 23 January 2026
Accepted 04 February 2026

***Corresponding Author; Taejung Park**

Tel: +82-2-901-8339
E-mail: tjpark@duksung.ac.kr

Ⅰ. 서 론

딥러닝 기반 음성 합성 기술은 자연스러운 발화 품질과 화자 유사도를 빠르게 달성하며 다양한 산업 분야에서 활용되고 있다[1],[7]. 그러나 이러한 기술은 동시에 보이스 피싱, 인증 우회, 사회 공학적 공격과 같은 악의적인 목적으로도 사용될 수 있어 심각한 보안 위협으로 작용하고 있다[1],[7]. 특히 음성 기반 인증 시스템이 널리 사용되는 환경에서는, 단순한 화자 검증(Automatic Speaker Verification, ASV)만으로는 합성 음성과 실제 음성을 구분하는 데 한계가 존재한다[7].

기존 음성 딥페이크 탐지 연구는 주로 시간 영역의 음성 파형(waveform) 또는 주파수 영역의 스펙트로그램(spectrogram) 중 하나의 표현에 집중해 왔다[3],[4]. 그러나 단일 표현 기반 접근법은 공격 유형이나 데이터 분포 변화에 취약하며, 학습 데이터에 포함되지 않은 unseen 공격에 대한 일반화 성능이 제한적이라는 문제가 지속적으로 보고되고 있다[1],[8].

이러한 한계를 극복하기 위해 최근에는 서로 다른 음성 표현한 후 이를 동시에 활용하는 멀티뷰 학습(Multi-view Learning) 접근이 주목받고 있다[1]. 그중 Multi-View Collaboraive Learning Network(MVCLN)[1]는 waveform과 spectrogram을 병렬적으로 학습하고, 두 뷰 간의 상호 보완적 특징을 결합함으로써 음성 딥페이크 탐지 성능을 향상시킨 대표적인 모델이다. 그러나 기존 MVCLN에서는 스펙트로그램 인코더로 Convolutional Neural Network(CNN) 기반 구조를 사용하고 있어, 주파수-시간 영역의 전역적 패턴을 충분히 반영하지 못한다는 한계가 있다.

본 연구에서는 이러한 문제를 해결하기 위해 기존 MVCLN 구조를 기반으로, 스펙트로그램 인코더를 Vision Transformer(ViT)[2] 기반 구조로 대체하고, waveform-spectrogram 간의 전역적 연간 관계 학습을 강화한 음성 딥페이크 탐지 모델을 설계하고자 한다. 이를 통해 CNN 기반 스펙트로그램 인코더가 포착하기 어려운 전역 주파수-시간 패턴을 효과적으로 학습하고, 다양한 딥페이크 공격 시나리오에서의 일반화 성능 향상을 목표로 한다.

Ⅱ. 음성 딥페이크 탐지 및 멀티뷰 학습 기반 접근

2-1 음성 딥페이크 탐지

음성 딥페이크 탐지는 실제 인간 음성과 합성 또는 변조된 음성을 구분하는 문제로 정의된다. 초기 연구에서는 Mel-Frequency Cepstral Coefficients(MFCC)[3], Constant-Q Cepstral Coefficients(CQCC)[4]와 같은 수작업 특징을 기반으로 Gaussian Mixture Model(GMM)[5]이나 Support Vector Machine(SVM)[6]과 같은 전통적인 통계 또는 인공신경망을 사용하지 않는 머신러닝 기반 분류기를 사용하였다.

이후 딥러닝 기술의 발전과 함께 Convolution Neural Network(CNN) 기반 스펙트로그램 분류 모델이 음성 딥페이크 탐지 분야에 본격적으로 적용되었으며, ASVspoof 챌린지[7],[8]를 계기로 다양한 딥페이크 탐지 모델이 제안되었다. 이러한 모델들은 학습 데이터 분포 내에서는 높은 탐지 성능을 보였으나, 학습 시 노출되지 않은 unseen 공격에 대해서는 성능 저하가 발생하는 문제가 지속적으로 보고되었다[8].

2-2 멀티뷰 학습 기반 탐지 모델

멀티뷰 학습(Multi-view Learning)은 동일한 데이터를 서로 다른 방식으로 표현한 후 표현된 결과들 역시 동일한 데이터에 기반하고 있다는 점에 착안하여 모델의 표현력(representative capacity)을 강화하는 방법이며 동일한 데이터를 다양한 측면에서 표현하기 위해서 널리 알려진 변환(transform)이 적용된다. 본 연구에서는 이러한 멀티뷰 학습 방식들 중 하나인 Multi-View Collaborative Learning Network(MVCLN)[1]을 개선하였는데, MVCLN은 입력된 음성 신호(sound)를 각각 시계열 정보인 음성 파형(waveform)과 주파수 영역으로 확장한 2차원 이미지 정보인 스펙트로그램(spectrogram)으로 표현한 후 각각 적절한 DNN(Deep Neural Network)으로 학습한 후 Cross Attention 메커니즘[9]을 통해 서로 다른 표현 결과물의 상호 관련성을 학습시키는 방식을 적용한다. MVCLN에서는 시계열 정보인 음성 파형 데이터는 1차적으로 1D-Convolution 레이어로 처리한 후 그 이후부터는 Transformer[9]를 사용해서 학습하고 이미지 정보인 Spectrogram은 CNN(2D-Convolution)을 이용해서 학습을 수행한다. Transformer 레이어와 CNN 레이어에서 출력되는 정보(representative vectors 또는 embedding vectors)는 Cross Attention 메커니즘으로 구성된 WSFM(Waveform-Spectrogram Fusion Module)[1]에서 상호 연관성을 학습함으로써 합성 음성인지, 실제 음성인지의 여부를 판단한다.

2-3 Vision Transformer

Vision Transformer(ViT)[2]는 이미지 패치를 토큰 단위로 처리하여 전역적인 관계를 학습할 수 있는 구조로, 이미지 처리에서 널리 사용되고 있는 CNN 모델을 대체 혹은 보완할 수 있는 유력한 모델로 그 활용도를 넓혀 나가고 있다. CNN 모델이 포유류의 망막 시신경 세포의 구조와 작용 원리를 모방한 수용장(receptive field)[10]에 기초하여 파악하고자 하는 픽셀과 가까운 주변 로컬 픽셀들 사이의 관계에 집중하는 특징을 가지고 있는데 이러한 특징 때문에 이미지 내에서 상대적으로 멀리 떨어져 있는 픽셀 혹은 사물들 사이의

관련성은 파악하기 어려운 구조적인 문제를 가지고 있다. 이에 비해서 ViT 기반 모델은 Attention 메커니즘에 기초하여 가까운 정보들과의 관련성뿐만 아니라 시간/공간 차원에서 멀리 떨어져 있는 정보들 사이의 관련성까지 파악할 수 있는 특성을 가지고 있다.

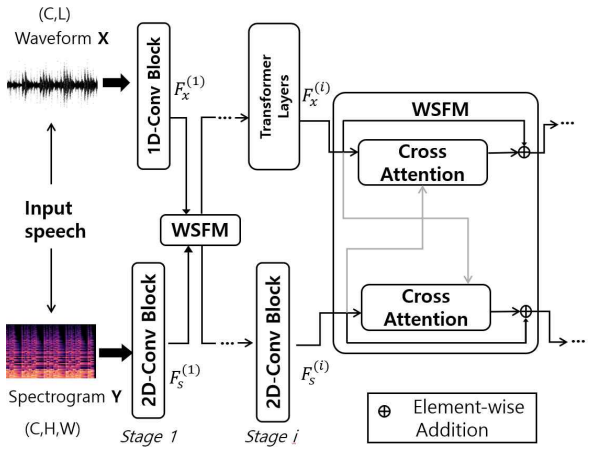


그림 1. 기존 MVCLN 구조도[1]
Fig. 1. Diagram of original MVCLN[1]

Ⅲ. 제안하는 ViT 기반 음성 딥페이크 탐지 기술

3-1 MVCLN의 한계 및 개선 아이디어

그림 1에서는 기존 MVCLN의 구조도를 제시한다. 2-2절에서 논의한 것처럼 MVCLN의 기본적인 아이디어는 원본 데이터(음성 신호)를 시간축(waveform)에서 Transformer로 학습하고(그림 1에서 위쪽 데이터 흐름), 다른 한편으로는 동일한 원본 데이터를 STFT(Short Time Fourier Transform)[11]로 변환하여 시간축과 주파수축으로 구성되는 2차원 평면 위에 주파수 성분의 크기를 색상값으로 표현하는 2차원 스펙트로그램 이미지로 나타낸 후에 CNN으로 이 스펙트로그램 이미지를 학습하게 된다(그림 1에서 아래쪽 데이터 흐름). 실제 적용에서 이러한 CNN 레이어는 ImageNet[12] 등 대규모 이미지 데이터셋을 이용해서 사전에 학습한 가중치를 적용한 네트워크를 이용하는 것이 일반적이다.

그러나 스펙트로그램 이미지의 경우, STFT의 원리를 고려해 본다면, 시간축과 주파수축 상에서 서로 가깝게 배치된 주변 정보들뿐만 아니라 상대적으로 멀리 떨어져 있는 정보들 사이에도 중요한 관련성이 내포되어 있을 가능성이 높다. 그러나 MVCLN에서 적용한 CNN은 일상적인 환경에서 개별 물체 및 물체들 사이의 상관 관계를 식별하기 위해 여러 픽셀들이 동일한 기하학적 단위에 속하는지의 여부(spatial cohesion[13])를 파악하고자 하는 목적으로 망막의 메커니즘을 모방한 수용장 단위에서 각 픽셀 주변 정보와의 관계를 계산하는데 최적화되어 있으나, 상대적으로 멀리 떨어진 픽셀 정보들의 상관 관계는 거의 고려되지 않는다는 점을 생각해 보면 스펙트로그램에 내포된 유용한 정보들을 충분히 포착하기 어렵다고 결론 내릴 수 있다.

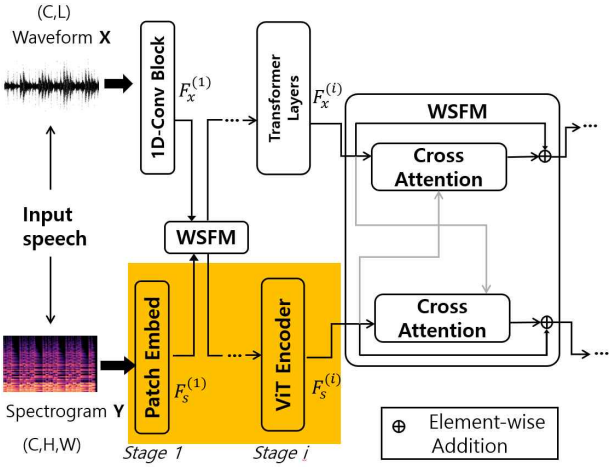


그림 2. 본 연구에서 제안하는 개선된 네트워크 구조(수정된 부분은 주황색 영역으로 제시)
Fig. 2. Improved network architecture proposed in this study (with the modified components highlighted in orange)

따라서 본 연구에서는 그림 2에서 제시한 것처럼, 스펙트로그램에서 서로 멀리 떨어져 있는 픽셀들 사이의 관계에 내포되어 있는 정보까지 활용할 수 있는 ViT를 이용해서 음성 딥페이크 탐지 성능을 보다 개선하는 방안을 제안한다.

3-2 데이터셋

본 연구에서는 공정한 비교를 위하여 MVCLN에서 적용된 데이터셋[14]-[16]을 참고하여 다양한 음성 딥페이크 공격 유형을 포함한 통합 데이터셋을 구성하였다. 데이터셋은 위조되지 않은 데이터(bonafide)와 페이크 데이터(spoof) 두 가지 클래스로 구성된다.

데이터는 학습(train), 검증(validation), 테스트(test) 데이터셋으로 분리하였으며, ASVspoof 챌린지 가이드라인[7],[16]에 따라 attack ID 기준으로 분할하였다. 테스트 데이터는 학습 과정에서 사용된 공격 유형으로 구성된 seen(test) 환경과, 학습 시 전혀 노출되지 않은 공격 유형으로 구성된 unseen(test_unseen) 환경으로 구분하였다. 이를 통해 모델의 분포 외 공격(unseen attack)에 대한 일반화 성능을 평가하고자 하였다.

모든 음성 파일은 16kHz로 리샘플링되었으며, 최대 6초 길이로 정규화하였다. 스펙트로그램은 Short-Time Fourier Transform(STFT)[11] 기반 로그 파워 스펙트럼을 사용하였다.

표 1. 데이터셋
Table 1. Dataset

Split	# Samples
Train	255,636
Validation	31,954
Test (Seen Attack)	31,955
Test-Unseen (Unseen Attack)	14,710
Total	**334,255**

3-3 전체 모델 구조

제안하는 모델은 대체적으로 MVCLN에서 제안한 음성 파형 처리 부분(waveform branch)과 스펙트로그램 처리 부분(spectrogram branch)으로 구성된 멀티뷰 구조를 따라 구성되지만 몇 가지 측면에서 개선이 적용되었다. 먼저 음성 파형 처리 부분에서는 MVCLN에서와 동일하게 1D-Convolution 기반 프론트엔드를 거친 후 여러 단계에 걸쳐 Transformer Encoder[9]를 사용하여 시간 영역 특징을 학습한다.

그러나 MVCLN에서 스펙트로그램 처리 네트워크에서 CNN(2D-Conv)을 이용하여 스펙트로그램을 학습한 것과는 달리, 앞서 논의한 CNN의 한계를 극복하기 위해서 스펙트로그램을 패치 단위로 분할한 후 Vision Transformer(ViT) Encoder[2]를 적용하여 주파수/시간 영역의 전역 패턴을 학습하도록 개선하였다.

스펙트로그램 입력은 STFT 기반 로그 파워 스펙트럼으로 생성되며, 입력 크기를 224×224로 정규화한 후 16×16 크기의 패치 단위로 분할한다. 이를 통해 총 196개의 패치 토큰이 생성되며, 각 패치는 선형 임베딩을 통해 768차원 토큰 벡터로 변환된다.

ViT 인코더는 총 12개의 Transformer Encoder 블록으로 구성되며, 각 블록은 12-head self-attention 구조를 따른다. 본 연구에서는 ImageNet[12]으로 사전 학습된 ViT 가중치를 초기값으로 사용한 후, 음성 스펙트로그램 데이터에 대해 미세 조정(fine-tuning)을 수행하였다. 이를 통해 CNN 기반 인코더가 포착하기 어려운 주파수-시간 영역의 전역적 패턴과 장기 종속성을 효과적으로 학습하도록 설계하였다.

각 뷰에서 추출된 waveform 특징과 spectrogram 특징은 이후 WSFM(Waveform-Spectrogram Fusion Module)을 통해 단계적으로 융합되며, 최종적으로 분류기에 전달된다.

3-4 WSFM 기반 뷰 간 융합

그림 1과 그림 2에서 제시한 것처럼 음성 파형 처리 부분과 스펙트로그램 처리 부분 부분에서 각각 추출된 특징은 WSFM(Waveform-Spectrogram Fusion Module)을 통해 상호 보완적으로 융합된다. WSFM은 각 뷰의 상호 연관 관계를 Cross Attention[9]으로 계산하고 전달함으로써, 음성 파형과 스펙트로그램 사이에서 내포하고 있는 유용한 정보를 포착하기 위한 협력 학습을 유도한다.

본 연구에서는 Explicit Attention 기반 WSFM과 Gated WSFM을 결합하여 이러한 상호 내포하고 있는 정보를 파악하는 융합 성능을 강화하였다.

Explicit Cross-Attention은 waveform 특징 F_x와 spectrogram 특징 F_s 간의 전역적 상관관계를 다음과 같이 계산한다.

$$Attention(Q_x, K_s, V_s) = softmax(\frac{Q_x K_s^T}{\sqrt{d}}) V_s \quad (1)$$

여기서 Q_x는 waveform 특징에서 생성된 query이며, K_s와 V_s는 spectrogram 특징에서 생성된 key와 value이다.

이후 Gated Attention 메커니즘을 통해 Explicit Attention 결과와 원래 waveform 특징 간의 기여도를 동적으로 조절한다. 게이팅 계수 G는 다음과 같이 정의된다.

$$G = \sigma(W_g[F_x; F_s]) \quad (2)$$

최종 융합 특징은 다음과 같이 계산된다.

$$F_{fused} = G \odot F_{explicit} + (1 - G) \odot F_x \quad (3)$$

이를 통해 Explicit Attetnion이 제공하는 전역적 상관 정보와 Gated Attention을 통한 선택적 정보 전달을 동시에 반영함으로써, 두 뷰 간 상호 보완적 특성을 효과적으로 결합하도록 설계하였다.

그림 2에서는 생략되었으나 그림 2에서 제시된 네트워크 구조는 여러 단계(stage)로 반복된다. 이러한 단계구성과 관련된 하이퍼파라미터(hyper parameter)는 공정한 비교를 위하여 [1]의 Implementation Details 절에서 논의한 대로 기존 MVCLN에서 실험한 내용과 동일하게 구성하였다.

3-5 분류기 및 학습 전략

그림 3에서는 마지막 단계에 해당하는 분류기(classifier)의 구조를 제시하고 있다. 특히, 전체 구조에서 동일한 음성 신호를 두 가지 다른 관점 또는 변환(즉, 음성 파형과 스펙트로그램)을 통해 서로 다른 뷰(표현)를 생성하였으나 두 가지 표현 모두 동일한 정보를 내포한다는 측면에서 대비 학습(contrastive learning)을 수행한다.

이 분류기는 MLP(Multi-Layered Perceptron)[17]로 구성되며 그림 2에서 제시한 여러 단계의 WSFM 네트워크를 거친 후 출력되는 벡터(feature vector 또는 embedding vector)인 F_x와 F_s를 입력으로 받아서 병합한 후 기존 방식과 동일하게 대비 학습(Contrastive learning)[18]을 수행한다. 이 학습 과정에서는 분류를 위한 기본적인 비용 함수인

Binary Cross Entropy(BCE) loss[19]와 함께 대비 학습을 위한 inter-view contrastive loss 및 intra-view supervised contrastive loss를 결합[1]하여 표현 간 정렬과 분리도를 동시에 향상시킨다.

모델 학습에는 Adam 옵티마이저를 사용하였으며, 초기 학습률은 1×10^{-4}, 배치 크기는 64, 총 50 epochs 동안 학습을 수행하였다. 학습률은 cosine annealing 스케줄러를 적용하여 점진적으로 감소시켰다.

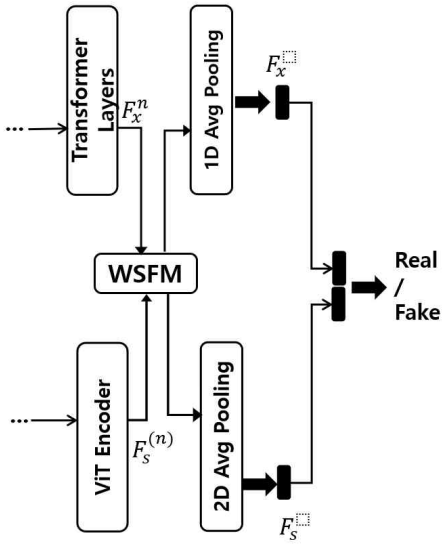


그림 3. 최종 분류기 네트워크 구조
Fig. 3. Final classifier network structure

Ⅳ. 실험 결과 분석

본 연구를 수행하는 시점에서 MVCLN 모델에 대한 소스코드는 공개되지 않아 [1]에서 제시된 네트워크 구조를 직접 구현한 후에 비교 연구를 수행하였다. MVCLN과 개선된 네트워크 구조 모두 PyTorch를 이용해서 구현하였으며 RTX A5000 GPU 8개, 64 코어 CPU, RAM 용량 126GB가 설치된 서버에서 학습과 테스트를 수행하였다.

본 연구에서는음성 딥페이크 탐지 성능을 평가하기 위해 Area Under the Receiver Operating Characteristic Curve(AUC)와 Equal Error Rate(EER)를 주요 평가 지표로 사용하였다. AUC는 분류기의 전체적인 판별 능력을 나타내는 지표로, 값이 클수록 실제 음성과 합성 음성을 잘 구분함을 의미하며, EER은 위양성률(False Acceptance Rate)과 위음성률(False Rejection Rate)이 동일해지는 지점의 오류율로, 값이 낮을수록 탐지 성능이 우수함을 의미한다.

표 2에서는 제안하는 모델과 기존 모델의 실험 결과를 제시한다. 실험 결과, seen 환경에서 제안하는 모델은 AUC 0.9999, EER 0.38%, Accuracy 99.63%를 기록하였다. 이 결과는 학습 데이터 분포 내에서는 거의 완벽에 가까운 탐지 성능을 의미한다. Unseen 환경에서는 AUC 0.9653, EER 8.56%, Accuracy 92.51%의 성능을 기록하였다. 이 수치는 기존 MVCLN에서 보고된 unseen 성능 대비 향상된 결과로, ViT 기반 스펙트로그램 인코더가 unseen 공격 유형에 대해 보다 일반화된 특징을 학습할 수 있음을 의미한다.

또한, CNN 기반 스펙트로그램 모델 및 waveform 기반 단일 뷰 모델들과의 비교에서도 제안하는 모델은 seen 및 unseen 환경 모두에서 일관되게 우수한 성능을 보였다. 이러한 결과는 waveform과 spectrogram을 동시에 활용하는 멀티뷰 구조와, ViT 기반 전역 주파수-시간 패턴 학습이 음성 딥페이크 탐지 성능 향상에 기여함을 정량적으로 뒷받침한다.

표 2. 제안 모델의 성능 평가 결과
Table 2. Performance of the proposed model

Model	MVCLN	Proposed Model
Seen AUC	0.953	0.999
Seen EER (%)	9.33	0.38
Unseen AUC	0.953*	0.965
Unseen EER (%)	9.85*	8.56

*presented as average values for comparison

전체 데이터셋 기준으로는 AUC 0.9890, EER 2.96%, Accuracy 97.38%의 성능을 기록하여, 다양한 공격 유형에 대해 균형 잡힌 탐지 성능을 유지함을 확인하였다.

Ⅴ. 결 론

본 연구에서는 기존 MVCLN 구조를 개선하기 위해 스펙트로그램 학습 과정에서 로컬 범위에서 주변 픽셀들과의 상관 관계에 집중하는 CNN의 한계를 극복하고자 하는 아이디어를 바탕으로 Vision Transformer(ViT) 기반 스펙트로그램 인코더를 적용한 음성 딥페이크 탐지 모델을 제안하였다. 또한 waveform과 spectrogram 간의 전역적 연관 관계를 효과적으로 학습하기 위해 개선된 Waveform-Spectrogram Fusion Module(WSFM)을 적용하였다.

실험 결과, 제안한 모델은 seen 환경에서 AUC 0.9999, EER 0.38%를 기록하여 학습 데이터 분포 내에서는 거의 완벽에 가까운 탐지 성능을 보였다. 또한 unseen 환경에서도 AUC 0.9653, EER 8.56%의 성능을 달성하여, 학습 시 노출되지 않은 공격 유형에 대해서도 기존 MVCLN 대비 향상된 일반화 성능을 확인하였다. 이러한 결과는 ViT 기반 스펙트로그램 인코더가 주파수-시간 영역의 전역적 패턴을 효과적으로 학습할 수 있음을 정량적으로 입증한다.

아울러 CNN 기반 스펙트로그램 모델 및 waveform 기반 단일 뷰 모델과의 비교 실험을 통해, 제안한 멀티뷰 구조가 다양한 공격 시나리오에서 보다 안정적인 탐지 성능을 제공

함을 확인하였다. 이는 서로 다른 음성 딥페이크 탐지 성능 향상에 효과적임을 시사한다.

향후 연구에서는 보다 다양한 딥페이크 생성 기법과 실시간 환경을 고려한 실험을 통해 제안한 구조의 확장 가능성과 실제 응용 환경에서의 활용 가능성을 검증하고자 한다.

참고문헌

[1] K. Zhang, Z. Hua, R. Lan, Y. Guo, Y. Zhang, and G. Xu, “Multi-View Collaborative Learning Network for Speech Deepfake Detection,” *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39, No. 1, April 2025. https://doi.org/10.1609/aaai.v39i1.32094

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, ... and N. Houlsby, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” in *9th International Conference on Learning Representations*, ICLR 2021, Virtual Event, Austria, May 3-7, 2021.

[3] A. H. Mansour, G. Z. A. Salh, and K. A. Mohammed, “Voice Recognition using Dynamic Time Warping and Mel-Frequency Cepstral Coefficients Algorithms,” *International Journal of Computer Applications*, Vol. 116, No. 2, pp. 34-41, April 2015.

[4] J. Yang, R. K. Das, and H. Li, “Extended Constant-Q Cepstral Coefficients for Detection of Spoofing Attacks,” in *Proceedings of the 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Honolulu: HI, pp. 1024-1029, 2018. https://doi.org/10.23919/APSIPA.2018.8659537

[5] Scikit-Learn. 2.1. Gaussian Mixture Models [Internet]. Available: https://scikit-learn.org/stable/modules/mixture.html.

[6] Scikit-Learn. 1.4. Support Vector Machines [Internet]. Available: https://scikit-learn.org/stable/modules/svm.html.

[7] A. Nautsch, X. Wang, N. Evans, T. H. Kinnunen, V. Vestman, M. Todisco, ... and K. A. Lee, “ASVspoof 2019: Spoofing Countermeasures for the Detection of Synthesized, Converted and Replayed Speech,” *IEEE Transactions on Biometrics, Behavior, and Identity Science*, Vol. 3, No. 2, pp. 252-265, April 2021. https://doi.org/10.1109/TBIOM.2021.3059479

[8] X. Wang and J. Yamagishi, “A Comparative Study on Recent Neural Spoofing Countermeasures for Synthetic Speech Detection,” in *Proceedings of the Interspeech 2021*, Brno: Czechia, pp. 4259-4263, 2021. https://doi.org/10.21437/Interspeech.2021-702

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, ... and I. Polosukhin, “Attention is all you need,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach: CA, pp. 6000-6010, 2017.

[10] L. Peng, X. Chen, J. Chen, W. Zhao, and X. Cao, “Understanding the Role of Receptive Field of Convolutional Neural Network for Cloud Detection in Landsat 8 OLI Imagery,” *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 60, pp. 1-17, 2022. https://doi.org/10.1109/TGRS.2022.3150083

[11] K. Gröchenig, *Foundations of Time-Frequency Analysis (Applied and Numerical Harmonic Analysis)*, Boston, MA: Birkhäuser Boston, 2001. https://doi.org/10.1007/978-1-4612-0003-1

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database,” in *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami: FL, pp. 248-255, June 2009. https://doi.org/10.1109/CVPR.2009.5206848

[13] L. R. Wanger, J. A. Ferwerda, and D. P. Greenberg, “Perceiving Spatial Relationships in Computer-Generated Images,” *IEEE Computer Graphics and Applications*, Vol. 12, No. 3, pp. 44-58, May 1992, https://doi.org/10.1109/38.135913

[14] Z. Wu, T. Kinnunen, N. Evans, and J. Yamagishi, Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2015) Database, University of Edinburgh, The Centre for Speech Technology Research (CSTR), 2015. https://doi.org/10.7488/DS/298

[15] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, The 2nd Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2017) Database, Version 2, University of Edinburgh, The Centre for Speech Technology Research (CSTR), 2018. https://doi.org/10.7488/DS/2332

[16] J. Yamagishi, M. Todisco, M. Sahidullah, H. Delgado, X. Wang, N. Evans, ... and A. Nautsch, ASVspoof 2019: The 3rd Automatic Speaker Verification Spoofing and Countermeasures Challenge Database, University of Edinburgh, The Centre for Speech Technology Research (CSTR), 2019. https://doi.org/10.7488/DS/2555

[17] Scikit-Learn. 1.17. Neural Network Models (Supervised) [Internet]. Available: https://scikit-learn.org/stable/modules/neural_networks_supervised.html.

[18] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A Simple Framework for Contrastive Learning of Visual Representations,” in *Proceedings of the 37th International Conference on Machine Learning*, pp. 1597-1607, 2020.

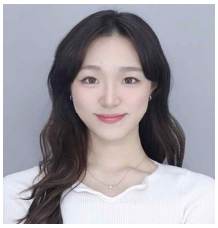
[19] PyTorch. BCELoss — PyTorch 2.9 Documentation [Internet]. Available: https://docs.pytorch.org/docs/stable/generated/torch.nn.BCELoss.html.

장혜준(Hyejun Jang)

2022년～현 재: 덕성여자대학교 사이버보안전공 학사 과정
※관심분야 : 인공지능 기반 시선 추적, AI 보안

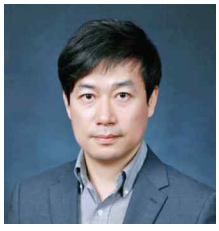
조서진(Seojin Cho)

2022년～현 재: 덕성여자대학교 사이버보안전공 학사 과정
※관심분야 : 딥페이크, AI 보안



김연수(Yeonsoo Kim)

2022년～현 재: 덕성여자대학교 사이버보안전공 학사 과정
※관심분야 : 디지털포렌식, AI 보안



박태정(Taejung Park)

1997년 : 서울대 전기공학부 (공학사)
1999년 : 서울대 전기공학부 대학원 (공학 석사, 반도체 물리 전공)
2006년 : 서울대 전기컴퓨터공학부 대학원 (공학박사, 컴퓨터 그래픽스 전공)
2006년～2013년: 고려대학교 연구교수
2013년～2017년: 덕성여자대학교 정보미디어대학 디지털미디어학과 조교수
2018년～2024년: 덕성여자대학교 공과대학 사이버보안/디지털소프트웨어공학부 부교수
2024년～현 재: 덕성여자대학교 공과대학 디지털소프트웨어공학부 교수
※관심분야 : 컴퓨터그래픽스, 인공지능, 수치해석, 3차원 모델링