Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.34, NO.3, Jun. 2024 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2024.34.3.439

언어 모델 기반 음성 특징 추출을 활용한 생성 음성 탐지*

김 승 민,[1†] 박 소 희,[2] 최 대 선[3‡]
[1,2,3]숭실대학교 (학생, 대학원생, 교수)

Voice Synthesis Detection Using Language Model-Based Speech Feature Extraction*

Seung-min Kim,[1†] So-hee Park,[2] Dae-seon Choi[3‡]
[1,2,3]Soongsil University (Student, Graduate student, Professor)

요 약

최근 음성 생성 기술의 급격한 발전으로, 텍스트만으로도 자연스러운 음성 합성이 가능해졌다. 이러한 발전은 타인의 음성을 생성하여 범죄에 이용하는 보이스피싱과 같은 악용 사례를 증가시키는 결과를 낳고 있다. 음성 생성 여부를 탐지하는 모델은 많이 개발되고 있으며, 일반적으로 음성의 특징을 추출하고 이러한 특징을 기반으로 음성 생성 여부를 탐지한다. 본 논문은 생성 음성으로 인한 악용 사례에 대응하기 위해 새로운 음성 특징 추출 모델을 제안한다. 오디오를 입력으로 받는 딥러닝 기반 오디오 코덱 모델과 사전 학습된 자연어 처리 모델인 BERT를 사용하여 새로운 음성 특징 추출 모델을 제안하였다. 본 논문이 제안한 음성 특징 추출 모델이 음성 탐지에 적합한지 확인하기 위해 추출된 특징을 활용하여 4가지 생성 음성 탐지 모델을 만들어 성능평가를 진행하였다. 성능 비교를 위해 기존 논문에서 제안한 Deepfeature 기반의 음성 탐지 모델 3개와 그 외 모델과 정확도 및 EER을 비교하였다. 제안한 모델은 88.08%로 기존 모델보다 높은 정확도와 11.79%의 낮은 EER을 보였다. 이를 통해 본 논문에서 제안한 음성 특징 추출 방법이 생성 음성과 실제 음성을 판별하는 효과적인 도구로 사용될 수 있음을 확인하였다.

ABSTRACT

Recent rapid advancements in voice generation technology have enabled the natural synthesis of voices using text alone. However, this progress has led to an increase in malicious activities, such as voice phishing (voishing), where generated voices are exploited for criminal purposes. Numerous models have been developed to detect the presence of synthesized voices, typically by extracting features from the voice and using these features to determine the likelihood of voice generation.This paper proposes a new model for extracting voice features to address misuse cases arising from generated voices. It utilizes a deep learning-based audio codec model and the pre-trained natural language processing model BERT to extract novel voice features. To assess the suitability of the proposed voice feature extraction model for voice detection, four generated voice detection models were created using the extracted features, and performance evaluations were conducted. For performance comparison, three voice detection models based on Deepfeature proposed in previous studies were evaluated against other models in terms of accuracy and EER. The model proposed in this paper achieved an accuracy of 88.08% and a low EER of 11.79%, outperforming the existing models. These results confirm that the voice feature extraction method introduced in this paper can be an effective tool for distinguishing between generated and real voices.

Keywords: BERT, Audio codec, Voice Features Extraction, Speech Synthesis, Generated voice detection

Received(01. 26. 2024), Modified(04. 26. 2024), Accepted(05. 20. 2024)

* 이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2021-0-00511, 엣지 AI 보안을 위한 Robust AI 및 분산 공격 탐지기술 개발).

† 주저자, kimsm0802@soongsil.ac.kr
‡ 교신저자, sunchoi@ssu.ac.kr(Corresponding author)

I. 서 론

최근 생성형 AI는 음성, 이미지, 자연어 처리 등 다양한 분야에서 활용되고 있다. 특히 음성 분야에서 생성형 AI는 활발한 연구가 이루어지고 있으며, 이는 텍스트를 음성으로 변환하는 TTS(Text-to-Speech)와 음성 변환 기술인 VC(Voice Conversion) 등이 대표적인 사례로 언급될 수 있다. 하지만 이러한 발전은 보이스피싱과 같은 악용 사례도 증가시키고 있다. 실제로 TTS나 VC와 같은 음성 합성 모델들은 다양한 응용 분야에서 활발한 사용을 보이며, 누구나 간편하게 접근하여 음성 생성 및 변환을 수행할 수 있다. 이러한 음성 기술의 악용을 예방하기 위해 생성 음성 탐지(detection) 및 검증기(verification)에 관한 연구도 활발히 진행되고 있다. 대표적으로 ASVspoof Challenge[1,2] 및 ADD Challenge[3]와 같은 활동을 통해 우수한 성능의 탐지 모델들이 개발되고 있다. 이러한 Challenge 들은 다양한 공격 환경과 방법으로 스푸핑 된 음성을 제공하며, 성능 측정을 위해 평가 지표를 설정한다. 음성의 특징을 추출하기 위해 어떤 방법을 사용할 것이며, 어떤 모델을 활용하여 분류를 수행할 것인지 등에 관한 다양한 연구를 진행한다.

본 논문은 위와 같은 생성 음성을 사용한 악용 사례에 대응하기 위해, ASVspoof에서 제공한 데이터셋을 사용하여 새로운 음성 특징 추출 모델(feature extraction)을 제안한다. 이를 위해, 자연어 처리 모델 BERT(Bidirectional Encoder Representations from Transformer)[4]와 오디오 코덱 모델인 Encodec[5]을 사용하였다. 제시된 음성 특징 추출 모델이 생성 음성과 실제 음성을 효과적으로 구별하는 특징을 추출할 수 있는지 검증하기 위해 4가지 생성 음성 탐지 모델을 만들어 실험을 진행하였다. 성능 평가를 위해 본 논문이 제안한 모델과 유사한 구조의 모델과 비교하였으며, 이에 더해 EER(Equal Error rate) 수치를 측정하였다. 본 논문에서는 음성 변환 과정 없이, 원시 오디오 데이터를 직접 사용하여 음성 특징을 추출하는 모델의 입력으로 사용하였다. 더불어, 음성 데이터도 텍스트와 마찬가지로 순차적인 정보가 중요하다는 사실을 강조하며, 음성과 텍스트를 유사한 접근 방식으로 다루었다. 이는 음성 데이터에 자연어 처리 기법을 효과적으로 확장하여 모델이 음성 데이터의 시간적 특성과 의미를 잘 이해할 수 있을 것으로 가정하였다.

본 논문에서 기여하는 바는 다음과 같다.

- 음성 데이터의 순차적인 특징을 효과적으로 추출하기 위해, 자연어 처리 모델 구조 기반의 새로운 음성 특징 추출기를 제안한다.
- 텍스트로 사전 학습된 특징 추출기를 활용하였을 때도 음성의 특징을 충분히 잘 추출 할 수 있으며, 이를 실제 음성과 생성 음성을 효과적으로 탐지할 수 있음을 증명하였다.
- 제안한 음성 특징 추출기를 이용하여 생성 음성을 감지한 결과, 기존의 생성 음성 탐지 모델보다 더 높은 정확도를 보였다.

II. 관련 연구

2.1 딥페이크 음성 탐지

오디오 딥페이크 탐지는 진짜 음성과 가짜 음성을 구별하는 작업이다. 딥페이크는 딥러닝 기술을 사용하여 사람들이 실제 음성으로 판별하기 어려운 자연스러운 가짜 오디오를 생성하는 것을 의미한다.

Fig 1에 따르면, 딥페이크 오디오 탐지는 주로 두 가지 유형의 방법으로 나눌 수 있다. 첫 번째 방법은 특징 추출기와 분류기로 이루어진 구조이다. 이 방법은 오디오에서 음성 특징을 추출하고, 추출된 음성을 기반으로 오디오가 진짜인지 가짜인지를 판별한다. 두 번째는 엔드투엔드 기반 탐지 방법으로, 최근 딥러닝 기술을 사용하여 발전되었다. 이 방법은 원시 오디오 형태에서 직접 작동하여 특성 추출과 분류를 한 번에 처리하며, 더 정교한 딥페이크 탐지가 가능해졌다.

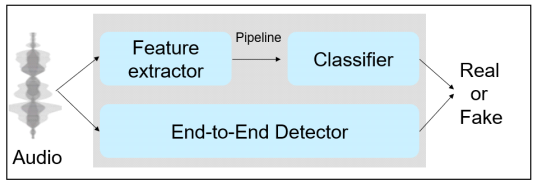

Fig. 1. Audio deepfake detection

2.2 음성 특징 추출

음성의 특징 추출은 음성 탐지 및 검증 과정에 있어 매우 중요한 단계이다. 실제 정상적인 음성과 생성된 음성들의 차이를 구분하기 위해서는 음성의 고유한 아티팩트(artifact)를 학습하고 특징을 잘 추출할 수 있어야 한다.

다음 논문[6]에 따르면 음성 특징 추출은 크게 Short-term spectral, Long-term spectral, prosodic features, Deep features 4가지 방법으로 구분할 수 있다.

2.2.1 Short-term Spectral

Short-term spectral은 주로 음성 신호에 단시간 푸리에 변환(STFT)을 적용하여 계산된다. 단시간 푸리에 변환은 수식 1과 같다.

$$X(t,\omega) = |X(t,\omega)|e^{j\phi(\omega)} \quad (1)$$

$|X(t,\omega)|$는 진폭 스펙트럼을 의미하고 $\phi(\omega)$는 프레임 t 및 주파수 ω에서의 위상 스펙트럼을 의미한다. Short-term spectral 기반 음성 특징 추출은 음성의 진폭 스펙트럼과 위상 스펙트럼을 사용하여 음성의 특징을 추출한다.

진폭 스펙트럼은 소리의 진폭을 주파수에 따라 표현한 것이며, 엔벨로프(envelopoe)와 같은 아티팩트가 나타난다. 이를 통해 생성 음성을 탐지할 수 있다. 진폭 스펙트럼 기반 특징 추출 방법에는 LMS(log magnitude spectrum)[7], MFCC(Mel-Frequency Cepstral Coefficients)[8] 등이 있다. 위상 스펙트럼은 생성 음성 모델에서 아티팩트로 최소 위상을 사용하는 방법 중 일부로, GD(group delay)[7], LPCC와 같은 방법들이 있다.

2.2.2 Long-term Spectral

Short-term spectral은 프레임별로 계산되기 때문에 음성의 시간적인 특성을 잘 추출하지 못한다. 이에 따라 음성 신호에서 장거리 정보를 추출하기 위해 시간-주파수 기반 특징 추출 방법인 Long-term spectral[9]이 제안되었다. 이는 신호의 주파수 변화를 시간에 따라 관찰하여 장기적인 특성을 파악할 수 있다. 이 방법은 낮은 주파수에서 발생하는 장거리 정보를 효과적으로 탐지할 수 있다. Long-term spectral은 주로 시간-주파수 분석 방법에 따라서 STFT, constant-Q transform(CQT), Hilbert transformer(HT) 및 wavelet transform(WT) 기반의 특징을 사용하며, 대표적인 방법으로 Modespec[10], CQCC[11], CFCC 등이 있다.

2.2.3 Prosodic Features

Prosodic features[12]는 음성 신호의 비구 간적인 정보를 나타내며, 음성의 강세, 억양, 말하기 속도, 리듬 등을 나타낸다. 이러한 특징은 spectral 특징과 달리 채널(channel) 효과에 민감하지 않으며, 생성 오디오 탐지의 성능을 향상하게 시키기 위해 spectral 특징과 함께 사용된다. 대표적인 방법으로는 기본 주파수(F0), 에너지 분포, 말하기 속도 등이 있다. 이러한 특성 때문에, prosodic 특징은 음성 인식, 감정 분석, 발화 인식 등 다양한 분야에서 사용된다.

2.2.4 Deep Features

Deep features는 DNN(Deep neural network)을 사용하여 음성을 학습하고 특징을 추출하는 방법이다. spectral 및 prosodic 기반 특징 추출은 사람이 직접 설계하고 추출한 것이기 때문에, 데이터의 복잡한 패턴을 충분히 파악하지 못한다. 반면에, DNN을 사용하여 학습된 특징은 입력 데이터의 복잡한 패턴을 더 잘 파악할 수 있다. 네트워크가 다층적으로 구성되어 각 레이어에서 데이터의 다양한 추상적 특성을 학습하고 깊게 쌓인 레이어들이 높은 수준의 추상화된 표현을 생성하기 때문이다. 이렇게 추출된 Deep features는 주로 학습 가능한 스펙트럼 피처(learnable spectral features), 지도형 임베딩 피처(supervised embedding features) 및 자기 지도형 임베딩 피처(self-supervised embedding features)로 구분된다. 대표적인 특징 추출 방법으로는, FastAudio, Wav2vec[13], XLS-R, HuBERT[14] 등이 있다.

2.3 ASVspoof Challenge

ASVspoof Challege는 오디오 딥페이크 감지 기술의 발전을 위한 대회로, 특히 스푸핑된 오디오로부터 ASV(Auto speaker Verification) 시 효과적으로 보호하는 것에 중점을 두고 다양한 기술들이 개발되고 있다. ASVspoof2015에서는 LA(Logical Access) 작업이 주로 다루어졌으며, 이는 합성 및 변환된 발화의 감지를 포함하고 있다. ASVspoof2017에서는 PA(Physical Access) 작업이 중심이 되었으며, 주로 리플레이 공격(replay attack)을 다루었다. ASVspoof2019[1]는 이전에 다루었던 PA와 LA 작업을 모두 포함하였으며, 이는 다양한 공격 환경

및 방법에서 ASV 시스템이 공격으로부터 견고해질 수 있게 목표로 하였다. 마지막으로 ASVspoof2021 [2]은 새롭게 DF(speech Deep fake) 작업을 추가하여 다양성을 확장하였다.

2.3.1 ASVspoof dataset

오디오 딥페이크 탐지 기술의 발전은 다양한 스푸핑 방법과 다양한 음향 조건을 고려하여 구성된 데이터셋에 의존하고 있다. 초기 연구에서는 ASV(Auto Speaker Verification)시스템에 대한 스푸핑 대응책을 개발하기 위해 다양한 스푸핑 데이터셋을 사용하였지만, 이러한 데이터셋은 특정 스푸핑 방법에만 의존하여 실험적으로 비교하기 어려웠다. 이 문제를 해결하기 다양한 TTS 및 VC 방법으로 구성된 표준 공개 스푸핑 데이터셋 SAS를 개발하였다. 이후 ASVspoof Challenge에서는 SAS 데이터셋을 활용하여 스푸핑 음성을 감지하는 연구가 진행되고 있다.

ASVspoof2019 데이터셋은 ASVspoof Challenge 2019에서 사용된 데이터셋으로, 리플레이 공격(replay attack), TTS-음성 합성 공격, VC-음성 변조 공격, 그리고 혼합 공격을 대상으로 하는 4개의 서브셋으로 이루어져 있다.

각 서브셋은 다양한 공격 유형을 대표하며, 다양한 환경에서 ASV 시스템의 견고성을 평가하는 데 사용된다. ASVspoof2021 데이터셋은 ASVspoof Challenge 2021에서 사용되는 최신 데이터셋으로 이전 데이터셋에서 DF(Deepfake) 공격이 추가되었다. 또한 데이터의 압축 효과를 고려하여 생성된 공격 음성이 포함되어 있어, 현실적인 환경에서의 음성 보안 시나리오를 반영하고 있다. 본 논문에서는 LA와 DF 같은 생성 음성에 관한 데이터셋에 중점을 두어, 생성 음성과 진짜 음성을 분류하는 작업을 수행하였다.

III. 제안 모델

본 논문에서는 BERT를 사용한 새로운 음성 특징 추출기를 제안한다. 본 논문에서는 음성 오디오 데이터를 텍스트와 유사한 방법으로 사용하기 위해 텍스트로 사전 학습된 BERT를 사용하였다. 더불어, BERT의 입력으로 음성 데이터를 사용하기 위해, 오디오 코덱 모델 Encodec을 BERT의 토크나이저로 사용하였다.

본 논문에서 제안하는 Encodec-BERT 모델은 기존의 HuBERT[14] 모델과는 명확하게 구분되는 기술적 접근을 제안한다. HuBERT는 음성의 음향적 특성을 자기 지도 학습 방식을 통해 분석하고 처리하여 음성 인식의 정확성을 향상하는 데 초점을 맞추고 있다. 반면, Encodec-BERT는 코덱 모델을 사용하여 음성 데이터의 음향적 특성을 더 세밀하게 표현하고 추출한 후, 이를 사전 학습된 BERT 모델의 입력으로 사용한다. 이러한 접근은 음성의 특성을 더욱 정확하게 분석하며, 특히 생성된 음성과 실제 음성을 구분하는 데 있어서 정확한 판별이 가능하다고 가정한다.

3.1 BERT

BERT[4]는 Transformer 아키텍처의 인코더를 여러 층 쌓아 올린 언어 모델이다. 주어진 단어 이전의 문맥만 고려하는 단방향 모델들과 달리, 주어진 단어의 좌우 양방향의 문맥을 모두 고려하여 해당 단어의 의미를 파악하는 자연어 처리 모델이다. 또한 문장 분류, 질문 응답, 번역 그리고 언어 모델링과 같은 특정 자연어 처리 작업에 대해 파인 튜닝을 통해 뛰어난 성능은 보인다. 기본 구조는 Transformer의 인코더를 쌓아 올린 구조이다.

본 논문에서는 음성의 특징 추출기로 BERT를 사용하였다. BERT를 사용한 음성 특징 추출은 오디오의 순차적인 정보를 양방향으로 고려하여 각 프레임에서의 음성 특징을 효과적으로 추출할 수 있을 것이라 가정하였다. 더불어, 적은 양의 데이터로도 높은 성능을 보일 것으로 예상하였다.

3.2 Encodec

소프트웨어에서 오디오 코덱이란 주어진 오디오 파일이나 스트리밍 미디어 오디오의 코딩 포맷에 따라 디지털 오디오 데이터를 압축하고 압축 해제하는 알고리즘을 구현한 컴퓨터를 의미한다. 이 알고리즘의 목적은 고품질 오디오 신호를 최소한의 비트로 효율적으로 표현하면서도 그 품질을 유지하는 것이다.

본 논문에서 사용한 Encodec[5]은 Meta에서 만든 코덱 모델이다. Fig 2과 같은 구조를 가지며, 고품질의 신경망 오디오 압축을 위해 설계되었다. 기본적으로 인코더와 디코더의 구조로 되어 있으며, 합성곱 블록 다음에는 시퀀스 모델링을 위한 두 개의 LSTM 레이어가 있다. 또한, 인코더와 디코더 사이에는 잔차 벡터 양자화(residual vector quantizer) 레이어가 있다.

잔차 벡터 양자화는 Fig 3과 같다. 오디오를 단일

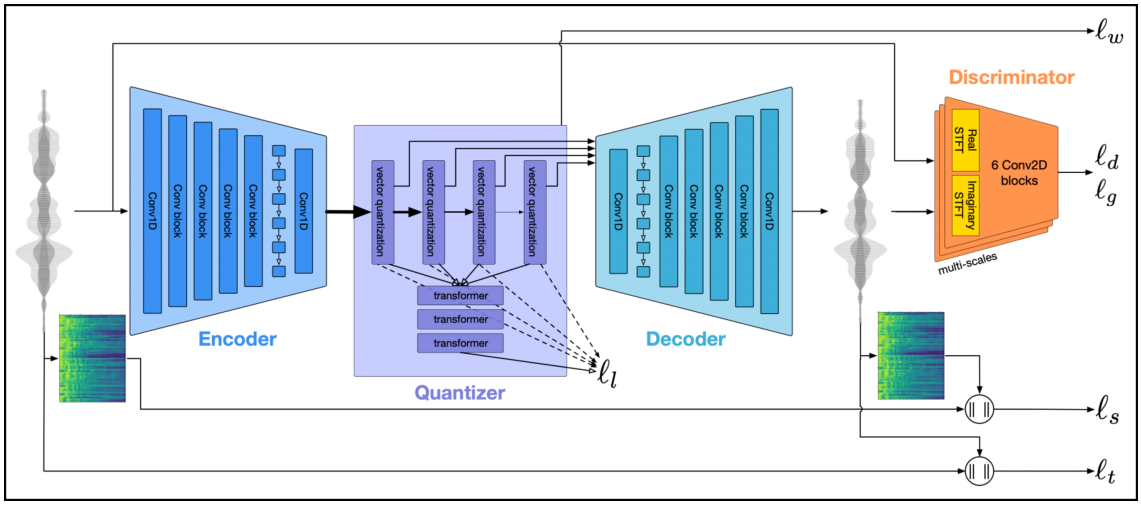

Fig. 2. The architecture of the Encodec model [5]

Residual Vector Quantization

Input: $y = enc(x)$ the output of the encoder, quantizers Q_i for $i = 1...N_q$
Output: the quantized $\hat{y}$
$\hat{y} \leftarrow 0.0$
residual $\leftarrow y$
for $i = 1$ to N_q **do**
  $\hat{y} \leftarrow \hat{y} + Q_i(residual)$
  residual $\leftarrow$ residual $-Q_i(residual)$
end for
return $\hat{y}$

Fig. 3. Residual Vector Quantization(RVQ) Algorithm [15]

코드북으로 표현하기에는 한계가 있으므로, 오디오 데이터를 효율적으로 표현하면서도 원래의 정보를 최대한 보존하기 위하여 잔차 벡터 양자화를 사용하였다.

3.3 BERT 기반 음성 특징 추출 모델 제안

본 논문은 Fig 4 와 같은 특징 추출기 모델을 제안하였다. 음성의 특징을 추출하기 위해 BERT를 사용하였으며, BERT의 토크나이저로 Encodec을 사용하였다. BERT의 입력은 토큰화를 통해 생성된 정수형의 시퀀스로 구성된다. 따라서, 논문 3.2절에서 언급한 바와 같이 Encodec의 인코더와 잔차 벡터 양자화 레이어를 사용하여 음성을 정수 형태로 바꾸어 BERT의 입력값으로 사용하였다. 양자화는 연속적인 데이터를 이산적인 값으로 표현하는 기술로, 음성의 파형을 특정 수준의 비트로 표현한 것이다. BERT 모델은 텍스트와 같게 이 이산적인 음성 정보를 처리할 수 있을 것이라 가정하였다. 양자화된 음성을 BERT가 텍스트와 유사한 방식으로 해석하고, 파인튜닝을 통해 음성을 학습하여 생성음성 과 실제 음성을 분류할 수 있는 중요한 특징을 추출할 것이라 가정하였다.

구체적인 실험에 앞서, 몇 가지 상황을 정의하였다. 본 논문의 제안 방법은 Encodec의 잔차 벡터 양자화 레이어를 거쳐 나온 음성의 임베딩 크기는 [2(양자화 레이어 개수), 양의 정수로 구성된 음성 특징] 형태로 출력되었다. 그러나 Bert의 입력은 최대 512 크기의 1차원 텐서이기 때문에, 전처리를 통해 [2, 512] 크기로 조정해 주었다. [2, 512] 형태도 BERT의 입력으로 적절하지 않았기 때문에, 첫 번째 값, 두 번째 값 그리고 평균값을 활용하여 실험을 진행하였다. 이전 논문의[16] 실험 결과를 통해 두 번째 양자화 레이어에서 나온 값을 사용할 때 탐지 성능이 높았기 때문에, 두 번째 값을 사용하였다.

본 논문에서는 BERT-Base 모델을 사용하였다. BERT-Base는 12개의 트랜스포머 인코더 레이어로 구성되며, 각 레이어는 768개의 히든 유닛과 12개의 셀프 어텐션 헤드를 포함한다. 본 논문에서는 생성 음성인지 아닌지 진위를 결정하는 분류 모델이기 때문에 모델의 입력은 '[CLS]'토큰으로 시작한다.

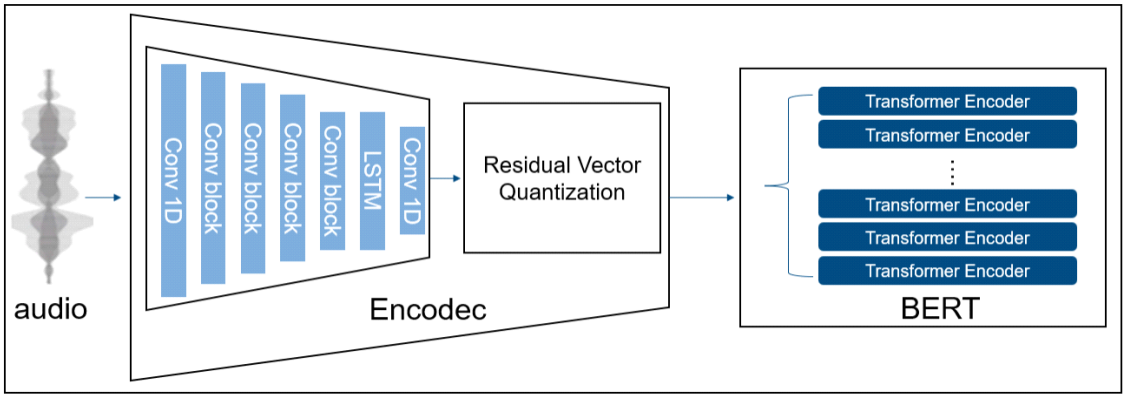

Fig. 4. The proposed feature extractor: Using Encodec as a Tokenizer to extract audio features with BERT

IV. 실험 및 실험 결과

4.1 실험 데이터

본 논문에서는 2.3절에서 언급한 ASVspoof2019 및 ASVspoof2021 데이터셋을 활용하여 실험을 진행하였다. 이 음성 데이터 셋은 생성 음성과 실제 음성으로 구성되어 있으며, 본 실험에서는 PA를 제외한 LA와 DF 데이터셋만을 사용하였다. 실험에 주된 목적은 생성 음성인지 실제 음성인지 구분하는 2가지 클래스에 대한 이진 분류이므로, 클래스 불균형 문제를 고려하는 것이 중요하다.

Table 1. ASVspoof2019 LA Dataset

	Bonifide	Spoof	Total
Train	2,580	2,580	5,160
Validation	2,548	2,548	5,096
Test	7,355	7,355	14,710

Table 2. ASVspoof2021 LA Dataset

	Bonifide	Spoof	Total
Train	10,784	10,784	21,568
Validation	2,704	2,704	5,408
Test	3,376	3,376	6,752

Table 3. ASVspoof2021 DF Dataset

	Bonifide	Spoof	Total
Train	11,824	11,824	23,648
Validation	2,548	2,548	5,096
Test	7,355	7,355	14,710

Challenge에서 제공하는 데이터셋은 한 클래스가 다른 클래스에 비해 8:1 정도로 상대적으로 더 많은 샘플을 가지고 있어, 이에 따라 샘플이 많은 클래스의 특징만을 주로 학습하는 클래스 불균형 문제가 발생할 가능성이 있다. 이를 대비하여 본 실험에서는 데이터셋을 1:1 비율로 전처리한 후 실험을 진행하였다. 전처리한 데이터셋의 수와 종류는 Table 1,2 그리고 Table 3과 같다.

4.2 실험 내용

논문 3.3절에서 제안한 음성 특징 추출 모델을 활용하여, 4.1에서 1:1 비율로 전처리 된 음성 데이터에 대한 실험을 진행하였다. 본 논문에 목적은 제안한 모델이 생성 음성과 실제 음성을 구분하는 데 있어서 효과적인 모델인지 확인하는 것이다.

실험에서는 제안한 음성 특징 추출 모델의 출력값을 Fig 5.와 같이 4가지 분류 모델의 입력값으로 사용하였다. 가장 단순한 방법으로는 FC-layer를 사용하였다. 이 모델에서는 제안한 모델로부터 추출된 특징값만으로도 생성 음성인지 실제 음성인지 분류되는지 확인하였다. 두 번째로는 Transformer 모델을 사용하였다. 자연어 처리 및 시퀀스 학습에서 높은 성능을 보이는 Transformer가 음성 데이터에도 효과적일 것이라 가정하였다. 세 번째로 이미지 분류에서 높은 성능을 보이는 ResNet-18[17] 모델을 사용하였다. 깊은 신경망 구조로 [512, 768] 크기의 고차원 음성 특징의 추상적인 특성을 잘 이해하고 학습할 수 있다고 가정하였다. 마지막으로 ResNet_2는 기존 ResNet-18를 구성하고 있는 각 레이어들의 입력과 출력 사이즈를 변형시킨 커스텀 모델이다.

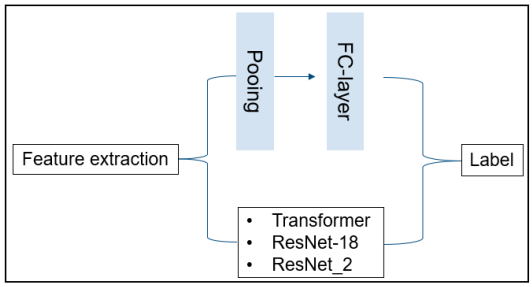

Fig. 5. The experimental methodology: Conducting experiments using four Classifier models

본 논문에서는 두 단계의 실험을 수행하였다. 먼저 ASVspoof2019 데이터셋을 사용하여 학습하고 평가하였다. 이러한 모델들을 ASSD[18]라는 음성 탐지 모델과 비교하였다. 다음 논문[18]에 제시된 내용에 따르면 ASSD 모델은 본 논문에서 제안한 모델과 유사한 구조를 갖추고 있으며, 오디오 코덱과 Transformer 기반의 신경망을 사용하고 있다.
따라서 본 연구의 제안 모델과의 성능 비교에 적합하다고 판단하였다. 더불어, ASSD 논문에서 성능평가로 사용된 3가지 모델에 대한 성능 비교도 추가하여 실험을 진행하였다.

두 번째로 ASVspoof2019 데이터셋을 기반으로 높은 성능의 분류기 모델 3개 정하여 실험을 진행하였다. ASVspoof2021데이터셋은 ASVspoof2019 데이터셋과 다르게 DF 영역이 추가되었다. DF와 LA 데이터셋을 개별적으로 학습시키고 평가하여, 어떤 종류의 데이터 영역에서 더 효과적으로 탐지하는지 비교 실험을 진행하였다. 또한, 본 논문은 생성된 음성을 탐지하는 것을 목표로 하므로, DF 데이터셋에서 기존 연구들과 비교하여 얼마나 효과적으로 음성을 탐지하는지 성능 비교를 진행하였다.

4.3 실험 결과

4.3.1 ASVspoof2019 데이터셋 실험 결과

ASVspoof2019 데이터셋을 사용하여 제안 모델과 ASSD 음성 탐지 모델과 비교한 결과는 Table 4와 같다. TSSDNet(Time-Domain Synthesis speech Detection Net)[19]와 CCT(compact Convolutional Transfomer)[20]은 기존 논문에서 ASSD와 함께 성능 비교를 위해 사용된 모델로 본 연구에서도 성능 비교를 위해 같이 사용하였다. 여기서, Class 0과 Class 1은 각각 실제 음성과 진짜 음성 클래스를 의미한다.

Class 0에 대한 정확도가 가장 높은 모델은 TSSDNet이지만, Class 1에 대한 정확도가 72.7%로 낮은 정확도를 보였고, 최종적으로 74.83%라는 가장 낮은 정확도를 보였다. CCT는 92.9%라는 가장 높은 정확도를 보였지만 Class 0과 Class 1의 정확도 차이가 너무 많이 나기 때문에 클래스 불균형 문제가 있다는 것을 확인할 수 있었다. 본 논문에서 제안한 4가지 모델과 ASSD 모델의 Class 0 정확도를 비교하면, ResNet_2와 Transformer가 약간 미세하지만, 더 높은 정확도를 보였다.
그러나, Class 1 정확도는 Transformer를 제외하면 더 높은 정확도를 보였다. 최종적으로 정확도가 가장 높은 모델은 ResNet_2였고 두 번째로는 FC-layer 모델이었다. 본 연구에서는 제안한 4가지 모델에 대해 추가로 EER(Equal Error rate)을 측정하였다. EER 결과에서도 FC-layer와 ResNet_2 모델이 낮은 수치를 보였다.

본 논문에서 제안한 특징 추출기는 3가지 분류 모델에서 기존 음성 탐지 모델보다 클래스별 및 전체 정확도는 높은 성능을 보였다. 그러나 클래스별 정확

Table 4. Experiment Results:ASVspoof2019 LA/ Comparison between paper[18]

Method	Class 0 Accuracy (%)	Class 1 Accuracy (%)	Balanced Accuracy (%)	Accuracy (%)	EER (%)
ASSD[18]	80.86	84.96	82.91	84.53	-
TSSDNet[19]	**92.32**	72.70	83.01	74.83	
CCT[20]	59.00	96.80	77.90	**92.90**	
FC-layer	80.84	**99.08**	89.96	87.88	17.09
ResNet-18	80.46	99.04	89.75	87.57	17.40
ResNet_2	81.27	98.69	**89.98**	88.08	15.98
Transformer	81.43	70.43	77.00	74.65	23.07

도의 차이를 통해 클래스 불균형 문제가 존재함을 알 수 있다. 이는 생성 음성과 실제 음성의 비율을 1:1로 전처리하는 과정에서 훈련 데이터셋, 검증 데이터셋 그리고 테스트 데이터셋의 비율이 한쪽으로 너무 치우쳐져, 훈련 데이터셋이 테스트 데이터셋에 비해 상대적으로 부족하므로 발생한 문제라 보인다.

4.3.2 ASVspoof2021 데이터셋 실험 결과

ASVspoof2021 DF, LA 데이터셋을 사용한 실험은 이전 ASVspoof2019 데이터셋을 사용한 실험에서 높은 성능을 보인 FC-layer와 ResNet_2 추가로 ResNet18을 사용하여 실험을 진행하였다. LA 데이터셋에 대해 3가지 모델 성능은 Fig 6과 같다. Class 0에 대한 정확도는 3가지 모델 모두 80%대를 보이지만 ASVspoof2019 데이터셋과 비교하였을 때 FC-layer가 약 4% 증가하여 가장 높은 84.52% 수치를 보였다. Class 1에 대한 정확도는 ResNet-18 모델이 95.34%로 가장 높은 정확도를 보였고, 전체 정확도에서는 Fc layer가 88.67%로 가장 높은 정확도와 10.91%라는 가장 낮은 EER을 보였다. 이는 예상과 달리 ASVspoof2021 LA 데이터셋에서 FC layer가 가장 높은 성능을 보였다.

DF 데이터셋에 대해 3가지 모델 성능은 Fig 7과 같다. Class 0에 대한 정확도는 LA 데이터셋을 사용한 실험과 동일하게 FC layer 모델이 86.56%로 가장 높은 정확도를 보였다. 하지만 Class 1에 대한 정확도는 ResNet_2 모델이 91.28%로 가장 높은 정확도를 보였다. 전체 정확도는 ResNet_2 모델이 87.02%로 가장 높은 정확도와 11.79%로 가장 낮은 EER을 보였다.

ASVspoof2021 LA와 DF에서는 FC-layer와 ResNet_2 모델이 각각 데이터셋에 따라 높은 성능을 보였다. 하지만, 모든 데이터셋에서 대부분 모델이 Class 0과 Class 1에 대한 정확도에서 불균형을 보였다. 그러나 Class 0에 대한 정확도가 85% 이상을 유지하고 있어, Class 1에 대한 정확도 향상을 통해 전체적인 모델의 성능을 높일 수 있을 것으로 판단하였다.

추가로, DF 데이터셋을 사용한 실험 결과를 바탕으로 기존 연구와 성능 비교한 결과는 Table 5와 같다. ASVspoof2021 Challenge의 DF 데이터셋에서 가장 높은 성능을 보인 모델은 Mel spectrogram, SinceNet 그리고 Raw를 음성 특징 추출기로 사용

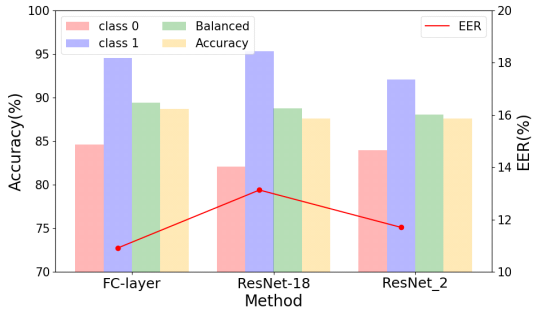

Fig. 6. Experiment Results: ASVspoof2021 LA

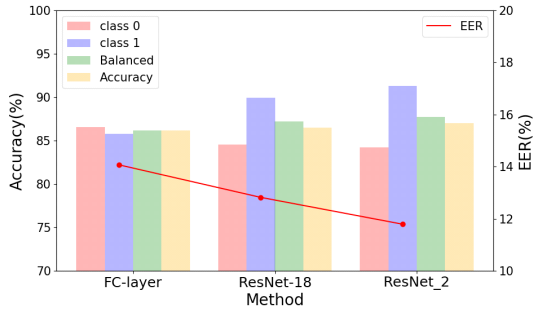

Fig. 7. Experiment Results:ASVspoof2021 DF

하고 LCNN과 ResNet을 분류기로 사용하여 15.64% EER을 보였다. 반면, 본 논문에서 제안한 Encodec-BERT를 음성 특징 추출기로 사용한 경우, 이전 연구들에 비해 낮은 EER 수치를 보였다. 특히, 세 가지 분류기 모델 중에서 ResNet_2 모델을 사용했을 때, 11.79%라는 가장 낮은 EER을 보였다. 이러한 실험 결과를 바탕으로 DF 데이터셋을 사용하여 생성된 음성을 탐지하는 데 있어 본 논문에서 제안한 모델이 기존 연구들보다 높은 성능을 보였다.

Table 5. Experiment Results:ASVspoof2021 DF/ Comparison between paper[6]

Method	Classifiers	EER (%)
Mel spec, SincNet, Raw	LCNN,ResNet	15.64
Fbank	ResNet, MLP	16.05
CQT	LCNN	18.30
Encodec-Bert	ResNet_2	**11.79**
	ResNet-18	12.82
	FC-layer	14.07

V. 결 론

본 논문은 생성 음성의 악용 사례에 대응하기 위해 사전 학습된 자연어 처리 모델인 BERT를 사용하여 음성 특징 추출 모델을 제안하였다. BERT에 알맞은 형태의 음성 값을 전달하기 위해, 오디오를 입력으로 받는 딥러닝 기반 오디오 코덱 모델 Encodec을 BERT의 토크나이저로 사용하였다. 본 논문이 제안한 음성 특징 추출 모델이 음성 탐지에 적합한지 확인하기 위해 추출된 특징을 활용하여 4가지 생성 음성 탐지 모델을 만들어 성능평가를 진행하였다.

ASVspoof2019 데이터셋을 사용하여 수행된 성능평가에서, 본 논문은 제안한 모델과 유사한 구조의 모델인 ASSD 음성 탐지 모델 외에 두 가지 모델과 성능을 비교하였다. Class 0에 대한 정확도는 TSSDNet 모델이 92.32% 수치로 가장 높았지만, class 0과 class 1에 대한 균형 정확도는 논문에서 제안한 FC-layer와 ResNet_2 모델이 각각 89.96%와 89.98%로 보다 높은 수치를 보였다. 또한 88.08% 정확도와 15.98% 낮은 EER 수치로 모든 모델 중 ResNet_2 모델이 가장 좋은 성능을 보였다.

ASVspoof2021 데이터셋은 ASVspoof2019 데이터셋을 기반으로 한 실험을 통해 본 논문에서 제시한 4가지 모델 중 성능이 높은 3가지를 선정하여 실험을 진행하였다. LA 샘플에서는 FC-layer모델이 Class 0과 Class 1의 균형 정확도가 89.42%로 가장 높았으며, 정확도 및 EER 수치도 각각 88.67%, 10.91%로 가장 좋은 성능을 보였다. DF 샘플에서는 Class 0과 Class 1의 균형 정확도가 87.74%로 가장 높았으며, 정확도 및 EER 수치도 각각 87.02%, 11.79%로 가장 좋은 성능을 보였다. 이 논문의 주요 목표는 생성된 음성을 탐지하는 것이며, DF 데이터셋을 사용하여 기존의 음성 탐지 방법들과 비교하였을 때 가장 낮은 EER을 보였다.

본 논문에서 제안한 음성 특징 추출 모델은 유사한 구조의 다른 모델들과의 비교에서 더 높은 성능을 보였다. Traonsformer를 사용한 분류기 모델을 제외하면 대부분 유사한 성능을 나타내어, 제안한 음성 특징 추출 모델이 어느 정도 수준의 효과를 가지고 있다고 판단된다. 따라서 자연어 처리 모델 BERT는 음성 데이터에 대해서도 효과적으로 특징을 추출할 수 있음을 보였다. 그러나 일부 모델들이 더 높은 정확도 및 낮은 EER 수치를 보여주고 있어, 더 나은 성능을 위한 추가적인 연구와 개발이 필요하다.

References

[1] Yamagishi, Junichi, et al. "Asvspoof 2019: Automatic speaker verification spoofing and countermeasures challenge evaluation plan." http://www.asvspoof.org/asvspoof2019/asvspoof2019evaluationplan. pdf, May, 2020

[2] DELGADO, Héctor, et al. Asvspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. arXiv preprint arXiv:2109.00535, 2021.

[3] Yi, Jiangyan, et al. "Add 2022: the first audio deep synthesis detection challenge." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.

[4] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv: 1810.04805, 2018.

[5] DÉFOSSEZ, Alexandre, et al. High fidelity neural audio compression. arXiv preprint arXiv:2210.13438, 2022.

[6] YI, Jiangyan, et al. Audio Deepfake Detection: A Survey. arXiv preprint arXiv:2308.14970, 2023.

[7] Xiao, Xiong, et al. "Spoofing speech detection using high dimensional magnitude and phase features: the NTU approach for ASVspoof 2015 challenge." Interspeech. 2015.

[8] Chen, Lian-Wu, Wu Guo, and Li-Rong Dai. "Speaker verification against synthetic speech." 2010 7th International Symposium on Chinese Spoken Language Processing. IEEE, 2010.

[9] Das, Rohan Kumar, Jichen Yang, and Haizhou Li. "Long range acoustic and deep features perspective on ASVspoof

2019." 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2019.

[10] M. Sahidullah, T. Kinnunen, and C. Hanilc¸i, "A comparisonof features for synthetic speech detection," in Proc. of INTER-SPEECH, 2015

[11] T. B. Patel and H. Patil, "Combining evidences from mel cepstral,cochlear filter cepstral and instantaneous frequency features fordetection of natural vs. spoofed speech," in Conference of Internaꟾtional Speech Communication Association, 2015

[12] T. K. A and H. L. B, "An overview of text-independent speakerrecognition: From features to supervectors," Speech Communicaꟾtion, vol. 52, no. 1, pp. 12–40, 2010

[13] Y. Xie, Z. Zhang, and Y. Yang, "Siamese network with wav2vecfeature for spoofing speech detection," in Interspeech, 2021

[14] X. Wang and J. Yamagishi, "Investigating self-supervised front ends for speech spoofing countermeasures," in The Speaker and Language Recognition Workshop, 2021

[15] Zeghidour, Neil, et al. "Soundstream: An end-to-end neural audio codec." IEEE/ACM Transactions on Audio, Speech, and Language Processing 30, 2021

[16] Kim Seung-min, et al., "Speech Generation Classifier Using BERT," Korea Institute of Information Security & Cryptology, September 2023.

[17] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016

[18] Yadav, Amit Kumar Singh, et al. "ASSD: Synthetic Speech Detection in the AAC Compressed Domain." ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023.

[19] G. Hua, A. B. J. Teoh, and H. Zhang, "Towards End-to-End Synthetic Speech Detection," IEEE Signal Processing Letters, vol. 28, pp. 1265–1269, Jun. 2021

[20] E. R. Bartusiak and E. J. Delp, "Synthesized Speech Detection Using Convoꟾlutional Transformer-Based Spectrogram Analysis," Proceedings of the IEEE Asilomar Conference on Signals, Systems, and Computers, pp. 1426–1430, Oct. 2021

〈저 자 소 개〉

김 승 민 (Seung-min Kim) 학생회원
2021년 3월~현재: 숭실대학교 소프트웨어학부 학사과정
〈관심분야〉 AI 보안, 생성 AI, 음성 처리

박 소 희 (So-hee Park) 학생회원
2018년 2월: 공주대학교 응용수학과 학사
2020년 2월: 공주대학교 융합과학과 석사
2020년 6월~2021년 9월: 한국교육학술정보원 전문원
2022년 2월~현재: 숭실대학교 소프트웨어학과 박사과정
〈관심분야〉 인증, 금융보안, 머신러닝, AI 보안, 적대적 공격 및 방어

최 대 선 (Dae-seon Choi) 종신회원
1995년 2월: 동국대학교 컴퓨터공학과 학사
1997년 2월: 포항공과대학교 컴퓨터공학과 석사
2009년 1월: 한국과학기술원 전산학과 박사
1997년 1월~1999년 6월: 현대정보기술 선임
1999년 7월~2015년 8월: 한국전자통신연구원 인증기술연구실 실장/책임연구원
2015년 9월~2020년 8월: 공주대학교 의료정보학과 부교수
2020년 9월~현재: 숭실대학교 소프트웨어학부 교수
2016년~현재: 정보보호학회 이사
〈관심분야〉 인증, 개인정보보호, AI 보안