딥보이스 탐지를 위한 딥러닝 모델 비교 분석

최길한, 유광운, 오찬영, 김용강*
국립 공주대학교

ygkim@kongju.ac.kr

A Comparative Analysis of Deep Learning Models for Deep Voice Detection

Gilhan Choi, Gwangun Yu, Chanyoung Oh, Yonggang Kim*
Kongju National University

요 약

인공지능이 고도화 됨에 따라 상대방의 목소리를 카피하는 딥보이스 생성형 AI 모델을 이용한 범죄 악용 사례가 증가하고 있다. 딥보이스를 탐지하기 위한 방법으로 다양한 음성 특성들이 활용될 수 있으며, 시간에 따라 변화하는 시계열 정보 또한 중요한 요소로 작용한다. 본 논문에서는 다양한 시계열 정보를 처리할 수 있는 딥러닝 모델들을 비교하여 분석한다.

Ⅰ. 서 론

최근 인공지능(AI) 기술의 발전과 함께 음성 합성 기술, 특히 딥보이스(Deep Voice) 기술이 급격히 발전하고 있다. 딥보이스는 딥러닝을 활용하여 인간의 목소리를 매우 사실적으로 모방하는 기술로, 사람의 음성 패턴을 학습하여 새로운 음성을 생성하는데 사용된다. 이러한 기술은 음성 비서, 가상 캐릭터, 자동화된 음성 응답 시스템 등 다양한 분야에서 활용될 수 있는 잠재력이 크다. 그러나 딥보이스 기술의 발전은 음성 보안과 개인정보 보호 측면에서 심각한 위험을 초래할 수 있다. 실제로 딥페이크(deepfake) 기술을 사용하여 특정 인물의 음성을 도용하거나, 잘못된 정보를 퍼뜨리는 등의 악용 사례가 발생하고 있으며, 이로 인해 개인 및 조직의 신뢰성에 큰 위협이 되고 있다. 특히 음성 인증 시스템을 공격하기 위한 수단으로 딥보이스 기술이 사용될 경우, 기존 보안 시스템을 무력화할 수 있는 위험성이 존재한다. 딥보이스의 이러한 잠재적인 위험성을 해결하기 위해, 딥러닝을 이용한 딥보이스 탐지 기술에 대한 연구가 활발히 진행되고 있다. 이에 따라 본 연구는 딥러닝 모델들의 성능을 평가함으로써 딥보이스 탐지 기술의 발전에 기여하고자 한다. 본 연구에서는 노이즈가 없는 짧은 길이의 음성 데이터에 대해 MFCC 전처리를 하여 서로 다른 딥러닝 모델의 성능을 비교 분석한다. 특히 real 음성과 fake 음성의 탐지 정확도를 높이기 위해 다양한 모델을 학습 및 테스트하여 성능을 분석하고, 각 모델의 강점과 한계를 평가하고자 한다.

Ⅱ. 딥보이스 탐지를 위한 실험 환경 구축

본 연구에서는 macOS 운영체제, Apple M1 Pro 칩, 16GB RAM, 1TB SSD 사양의 MacBook 환경에서 실험을 진행하였다. 데이터 분석 및 머신러닝 학습을 위해 Kaggle 플랫폼을 사용하였으며, 실험에 필요한 음성 데이터는 The Fake-or-Real (FoR) Dataset 에서 확보하였다. FoR 데이터 세트는 for-original, for-norm, for-2sec, for-rerec 의 네 가지 버전으로 구성되어 있으며, 본 연구에서는 모든 음성 파일이 2 초로 잘린 for-2sec 버전을 사용하였다. for-2sec 데이터 세트는 모든 음성 파일이 2 초로 고정되어 있어, 모델이 일정한 시간 범위 내에서 음성 특징을 학습하고 비교하는 데 유리하다. 이는 가변 길이 음성 데이터를 처리할 때 발생하는 복잡성을 줄여주며, 2 초로 제한된 데이터는 연산 자원을 절약하고 데이터 처리 시간을 단축시켜 모델의 학습 및 테스트 과정이 보다 효율적으로 이루어질 수 있다. 또한, 고정된 길이의 데이터는 모델 간 성능 비교를 일관성 있게 수행할 수 있게 해준다. 더 나아가, 실제 음성 인증 시스템이나 음성 명령 시스템과 같이 짧은 길이의 음성 데이터를 다루는 응용 환경을 시뮬레이션하는 데 적합하기에 for-2sec 데이터 세트를 사용하였다.

Ⅲ. MFCC 를 활용한 데이터 전처리

본 연구에서는 Mel-Frequency Cepstral Coefficients (MFCC)를 사용하여 for-2sec 데이터 세트를 전처리하였다. MFCC 알고리즘은 다음의 6 단계로 구성된다. 첫 번째 단계는 Pre-emphasis 로, 고주파 성분을 강조하기 위해 신호를 필터로 처리하는 과정이다. 이 과정은 다음과 같은 수식(1)에 의해 정의된다.

$$Y[n] = X[n] - 0.95 \times X[n-1] \quad (1)$$

여기서 X[n]은 입력 신호, Y[n]은 출력 신호를 의미한다. 두 번째 단계는 Framing 으로, 음성 신호를 20~40 밀리초 구간으로 나누어 처리한다. 일반적으로 100 개의 샘플이 중첩되고, 각 프레임은 N=256N = 256N=256 샘플로 구성된다. 세 번째는 Hamming Window 로, 각 프레임에 Hamming Window 를 적용하여 인접 주파수 성분을 결합한다. Hamming Window 는 다음과 같은 수식(2)으로 정의된다.

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \quad (2)$$

여기서 N 은 각 프레임의 샘플 수이다. 네 번째는 Fast Fourier Transform (FFT)을 통해 시간 도메인의 신호를 주파수 도메인으로 변환하여 주파수 스펙트럼을 구하는 과정이다. 이는 다음 수식(3)에 의해 표현된다.

$$Y(w) = FFT[X(t)] \quad (3)$$

다섯 번째 단계는 Mel Filterbank 로, Mel Scale 을 사용하여 인간 청각 시스템에 맞게 주파수 스펙트럼을 추출한다. 주어진 주파수 f 에 대해 멜 변환은 다음 수식(4)으로 정의된다.

$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad (4)$$

마지막 단계는 Discrete Cosine Transform(DCT)으로, 로그 멜 스펙트럼을 시간 도메인으로 변환하여 최종적으로 MFCC 계수를 추출하는 과정이다. 이 결과는 음향 벡터로 변환되어 모델의 입력으로 사용된다.[1]

Ⅳ. 딥러닝 모델의 성능 평가: LSTM, GRU, 1D-CNN, Transformer

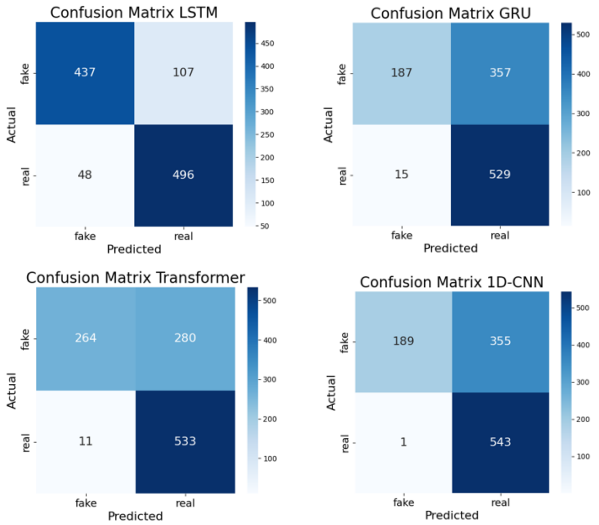

그림 1. 각 모델 별 Confusion Matrix

그림 1 를 보면 LSTM 모델과 GRU 모델의 혼돈행렬(Confusion Matrix)을 볼 수 있다. 혼돈행렬은 모델의 예측을 보다 보기 쉽게 정리 한 표로써 True Positive(TP), False Negative(FN), False Positive(FP), True Negative(TN)로 정리된다. LSTM 모델 같은 경우는 437 개의 데이터를 TN, 496 개의 데이터를 TP 로 올바르게 예측했다. 반면에 107 개의 데이터를 FP, 48 개의 데이터를 FN 로 잘못 예측했다. GRU 모델 같은 경우는 187 개의 데이터를 TN , 529 개의 데이터를 TP 로 올바르게 예측했다. 반면에 357 개의 데이터를 FP 로 15 개의 데이터를 FN 로 잘못 예측했다. Transformer 모델 같은 경우는 264 개의 데이터를 TN, 533 개의 데이터를 TP 로 올바르게 예측했다. 반면에 280 개의 데이터를 FP 로 11 개의 데이터를 FN 로 잘못 예측했다. 1D-cnn 모델 같은 경우 189 개의 데이터를 TN , 543 개의 데이터를 TP 로 올바르게 예측했다. 반면에 355 개의 데이터를 FP 로 1 개의 데이터를 FN 로 잘못 예측했다. 그림 1 를 종합해 보면 위 4 가지 모델 전부 Real 데이터를 잘 맞추는 경향이 있지만 Fake 구분에 있어서는 상대적으로 약한 모습을 보이고 있다.

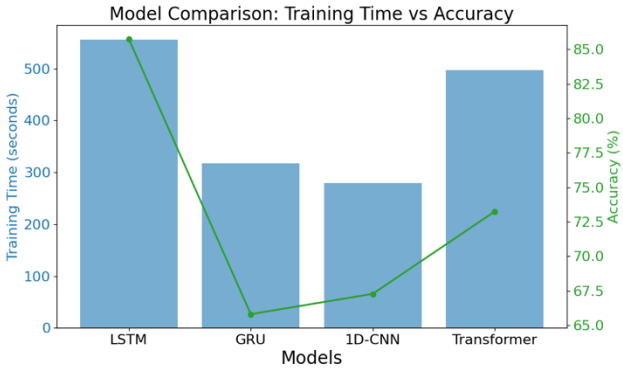

그림 2. Model Comparison

그림 2 를 보면 각 모델 별 학습 시간 및 정확도를 볼 수 있다. 시계열 데이터를 처리함에 있어서 장기적 의존성을 학습할 수 있도록 고안된 모델인 LSTM 이 가장 좋은 성능을 보여준다. 반면에 LSTM 의 장기적 의존성 학습을 단순화 한 모델인 GRU 같은 모델은 정확도가 많이 떨어지는 모습을 보여준다. 1D-CNN 같은 경우는 병렬처리가 가능하다는 점으로 인해 가장 빠른 학습 속도를 보이고 있지만 단기적인 데이터를 학습하는데 유리하게 설계된 모델 특성상 정확도는 다소 떨어짐을 보여준다. Transformer 모델 같은 경우에는 긴 시퀀스 내에서 장거리 관계를 학습에 효과적인 특성상 본 연구에서 이용된 2 초 짜리 음성 데이터에는 적합하지 않았음에도 2 번째로 좋은 정확도를 보여준다. 이에 상대적으로 짧은 시퀀스를 가진 시계열 데이터에는 LSTM 이이 적합하다는 것을 보여준다.

Ⅴ. 결론

본 연구에서는 딥보이스 기술을 탐지하기 위해 MFCC 전처리를 적용한 딥러닝 모델 LSTM, GRU, 1D-CNN, Transformer 의 성능을 비교 분석하였다. 비교적 적은 양의 시계열 데이터를 처리 하는 것은 LSTM 이 유리함을 확인할 수 있다. 향후 연구에서는 노이즈가 가미된 음성 데이터셋을 분석하여 Diffusion Model 을 활용한 노이즈 생성 및 음성 데이터 합성 방법을 도입할 예정이다. 이를 통해 노이즈가 포함된 환경에서 fake 음성을 보다 정확하게 구별하는 강건한 알고리즘을 개발하고, 딥보이스 기술의 보안 위협에 더욱 효과적으로 대응할 수 있는 방안을 모색할 것이다.

ACKNOWLEDGMENT

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. RS-2022-00166739). 또한 2024 년 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업 지원을 받아 수행되었음 (2024-0-00073)

References

[1] Lindasalwa Muda, Mumtaj Begam, I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", *arXiv:1003.4083*, 2010. (https://doi.org/10.48550/arXiv.1003.4083)