딥보이스 학습 방해를 위한 청각 마스킹 효과 기반 적대적 노이즈 활용 기법 연구

이강봉*, 조영호(교신저자)**
*국방대학교 국방관리대학원 국방과학학부 사이버컴퓨터공학과 석사과정
**국방대학교 국방관리대학원 국방과학학부 사이버컴퓨터공학과 교수
e-mail:lgb4521@naver.com, younghocho@korea.kr

A Study on Adversarial Noise Technique based on Auditory Masking Effect for Disrupting Deep Voice Learning

Gangbong Lee*, Youngho Cho**
*Master's Course, Dept. of Cyber Security and Computer Engineering, Korea National Defense University
**Professor, Dept. of Cyber Security and Computer Engineering, Korea National Defense University

요 약

딥보이스(Deep Voice)는 딥러닝 기술을 활용하여 사람의 목소리를 학습해 복제 또는 변환하는 기술이다. 최근 딥보이스 기술을 악용한 음성데이터의 도용·사기 사례가 지속 발생하고 있다. 이에 대응하기 위해 기존에 딥보이스 탐지기법에 대한 연구들이 수행되었으나, 이는 사후 식별방안으로 탐지 기술을 우회하여 음성데이터 악용이 가능하다는 제한점이 있다. 따라서, 본 논문에서는 사람의 청각 특성을 고려한 청각 마스킹 효과 기반의 적대적 노이즈를 활용하여 딥보이스 학습을 방해하는 사전 예방기법을 제안한다. 제안기법은 화자(Speaker)의 발화 구간에서만 마스킹 임계값 이하의 적대적 노이즈를 실시간으로 합성하고, 진폭변조(AM)와 주파수변조(FM)를 결합한 적대적 노이즈를 삽입함으로써, 사람의 청취에는 영향을 주지 않으면서도 음성인식(STT), 음성합성(TTS) 모델에서의 학습과 정렬 과정에서 오류를 유발하는 기법을 제시한다. 초도 실험 결과, 제안기법은 사람의 청취 품질을 유지하면서도 STT 및 TTS모델의 오류를 유발해 음성데이터 악용에 대한 방어 가능성을 확인하였다.

1. 서론

최근 딥러닝 기반의 신경망 모델의 진화와 적용으로 특정인의 목소리를 딥러닝 기술로 학습시켜 복제, 변환하는 기술인 딥보이스(Deep Voice) 기술이 발전하고 있다. 현재 딥보이스를 통해 누구나 쉽고 정교하게 목소리를 복제할 수 있는 수준에 이르렀다[1]. 지인이나 공공기관의 목소리를 사칭한 보이스피싱의 범죄가 현실화되고 있으며, 실제로 20초 가량의 음성 데이터만으로도 가족 음성을 모방하여 금전을 요구하는 상황을 만들 수 있다. 이에 따라 2025년 정부에서는 보이스피싱 근절 종합대책을 발표하는 등 딥보이스 기술이 실생활에 악용되는 것을 막기 위해 전력을 기울이고 있다.

이러한 상황에서 음성데이터 자체를 보호할 수 있는 기술의 필요성이 점점 커지고 있다. 특히, 음성 데이터는 강연, 발표, 유투브 등 대중에 음성데이터가 노출되는 상황에서 누구나 손쉽게 수집할 수 있는 표적이 될 수 있어 딥보이스를 이용한 범죄의 위험성이 매우 크다.

이를 대응하기 위해, 딥보이스 공격이 발생한 이후에 음성조작의 여부를 판별하는 탐지기법에 대한 연구들이 활발히 이루어졌으나, 이를 우회한 공격행위가 가능함에 따라 예방 차원에서 음성을 보호하는 방어기법이 절실하다.

따라서, 본 논문은 사람의 청각 특성을 고려해 음성을 숨기는 청각 마스킹 효과(Psychophysical Masking Effect)를 이용하여 사람의 귀에는 들리지 않지만 딥보이스나 음성인식 모델에는 학습에 방해를 유발하는 적대적 노이즈(Adversarial Noise)를 활용하여 딥보이스의 학습을 방해하는 선제적 예방기법을 제안한다.

2. 배경지식 및 기존연구

2.1 청각 마스킹 효과(Psychophysical Masking Effect)

청각 마스킹은 하나의 소리가 다른 소리를 들리지 않게 만드는 현상으로, 시간·주파수·세기 차이에 의해 특정 주파수 성분이 사람 귀의 임계치 아래에 들어가면 들리지 않는다는 개념이다[2]. 제안기법은 화자(Speaker)의 발화 구간에서만 시간·주파수 기반의 마스킹 임계값을 계산한 뒤, 그 임계값 이하의 노이즈를 삽입하여 청취자가 거의 지각하지 못하도록 설계한다.

2.2 적대적 예제 관련 기술(Adversarial Example Techniques)

적대적 예제(Adversarial Example)는 사람이 거의 감

지하지 못할 작은 섭동(Perturbation)이 원래 데이터에 더해 기계가 잘못된 예측을 유도하는 입력이다[3]. 본 연구에서는 오디오 신호에 적대적 노이즈를 삽입하여 모델의 예측을 교란하는 Evasion Attack의 형태로, STT(Speech to Text), TTS(Text to Speech)모델의 오류를 유발하는 기법을 제시한다[4].

2.3 기존 연구

한승우 등[5]은 음성 딥페이크 탐지시스템 개발을 위해 Mel Spectrogram, MFCC(Mel-Frequency Cepstral Coefficient) 특징을 분석하여 원본 음성과 조작된 음성을 구별하는 분류 모델을 연구했다. 하지만 이러한 탐지 및 식별 방안은 공격자에 의해 음성 데이터가 이미 확보된 상황에서 이루어지기 때문에 이를 우회하는 다양한 공격시도가 있을 수 있어 사전에 음성 데이터를 보호할 수 있는 예방 기술이 필수적이다.

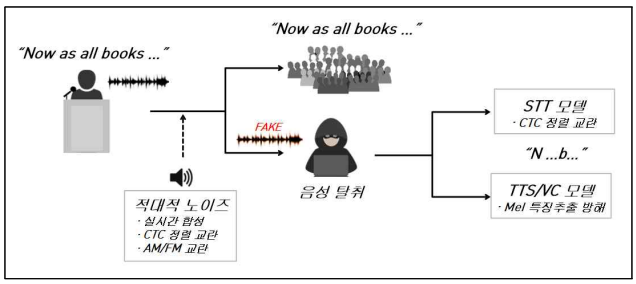
Fei 등[6]의 VocalCrypt에서는 인간 청각 특성을 고려한 청각 마스킹 효과를 활용하여 사람 귀에는 잘 들리지 않게 가짜 음색(Pseudo Timbre)를 삽입하여 음성 변환(Voice Conversion) 시스템이 원본 화자의 음색 정보 제대로 추출하지 못해 음성 복제 성능을 저하시키는 연구를 수행했다. 또한 실시간으로 노이즈를 생성하고 삽입하기 위해 Discrete Wavelet Transform(DWT), Quantization Index Modulation (QIM) 방식을 적용하였다. DWT를 통해 음성을 시간-주파수 대역으로 분해하여 특정 구간에만 노이즈를 삽입할 수 있도록 하고, QIM을 통해 삽입된 신호를 양자화 기반으로 임베딩해 압축, 재녹음에도 강건성을 높였다.

Zeng 등[7]은 Amplitude Modulation(AM)과 Frequency Modulation(FM) 성분이 음성인식 과정에서 핵심적인 역할을 한다는 것을 증명하며, 음성인식 성능 향상을 위한 방안을 제시했다. AM 성분이 발화 음성의 구분(문장 및 단어 구분)에 기여하며, FM 성분은 발음이나 화자 인식에 역할을 하는 것을 실험적으로 증명하였다. 이는 AM과 FM 성분을 인위적으로 교란할 경우 음성인식 모델이 정상적인 특징 추출에 실패하여 음성 인식율을 감소시킬 수 있음을 의미한다.

3. 제안기법

3.1 아이디어 및 설계

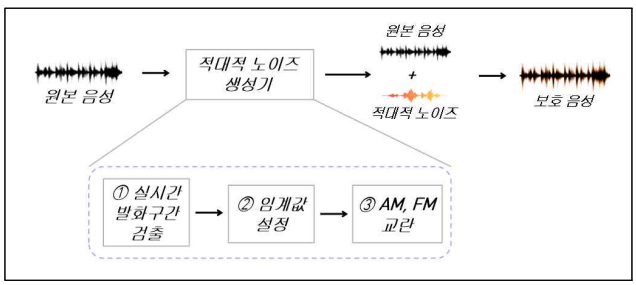
본 연구에서의 공격 상황과 제안기법을 활용한 방어전략을 나타내는 시나리오는 그림 1과 같다. 화자(Speaker)가 공개된 장소에서 발화 시 공격자가 화자의 음성 데이터를 탈취하는 행위를 예방하기 위해, 사람의 청각으로는 인지하지 못 하지만 기계에 입력 시 오류를 일으키는 적대적 노이즈를 생성한다. 적대적 노이즈는 청각 마스킹 효과를 기반으로 생성되며, 음색 차원에 국한되지 않고 CTC(Connectionist Temporal Classification) 정렬 과정 교란을 통해 STT 모델에 오류를 유발한다. 또한, AM, FM 성분 변조를 통해 TTS 모델의 Mel 특징 추출 안정성에 영향을 주어 학습을 방해한다.



[그림 1] 전체 시나리오

기존 연구[6]는 DWT, QIM으로 음성파동을 세분화하여 미세한 워터마크를 숨기는 방식이나, 본 연구에서는 AM, FM을 통해 음성 파형을 직접적으로 변형하고, Mel 특징추출과 CTC 정렬을 불안정하게 만드는 효과에 초점을 둔다.

3.2 적대적 노이즈 적용단계



[그림 2] 제안기법 개념도

제안기법에서 적대적 노이즈가 원본 음성에 적용되는 과정은 그림 2와 같이 3단계로 진행된다.

① 실시간 발화구간 검출 단계: 적대적 노이즈는 불필요한 잡음을 최소화하고 청취 피로감을 줄이기 위해 화자의 발화 구간에서만 활성화된다. 이를 위해 음성활동탐지(Voice Activity Detection)를 통해 실시간으로 프레임별(10ms) 발화 여부를 판별하고, 비발화 구간에서는 적대적 노이즈를 차단한다.

② 마스킹 임계값 설정 단계: 사람의 청각 특성을 고려한 Bark Scale을 기반으로 25개의 주파수 대역을 나누고, 특히 주요 음성 정보가 집중되어있는 300 ~ 2,200Hz 구간에 적대적 노이즈를 삽입하도록 설정한다[7]. 이때 삽입되는 신호의 대역별 출력 에너지는 각 대역의 마스킹 임계값의 70 ~ 80% 수준으로 제한되어 청중은 인지하지 못하지만 모델의 음성처리 과정에서 교란을 유발한다.

③ AM, FM 교란 단계: 적대적 노이즈가 단순 백색소음이나 잡음으로 인식되어 쉽게 분리·제거되지 않도록 음성 구조와 유사한 패턴으로 설계한다. 이를 위해 미세한 진폭변조(AM)와 주파수 변조(FM)를 적용하였으며, AM은 진폭의 변화를 주어 발화 리듬이나 강세와 같은 음성의 시간적 흐름을 교란시켜 음성인식에 필요한 음성정보 추출을 방해한

다. FM은 발음 구분에 중요한 모음 성분과 음높이 등의 주파수 성분을 흔들어 모델의 발음 판별을 어렵게 만든다. 이러한 AM, FM 교란은 사람이 거의 감지하지 못하는 범위에서 모델의 특징 추출 안정성을 약화시켜 인식 오류를 유발하는 것을 목표로 한다.

4. 초도 실험 결과

4.1 실험 목적 및 방법

본 실험은 음성이 노출된 환경에서 STT 및 TTS 모델 대상 악의적 학습을 방해하기 위해 적대적 노이즈를 생성/합성하여, 청각품질을 유지한 채 음성처리 모델의 학습의 교란 여부 검증을 목표로 한다. 실험 데이터는 공개 TTS 데이터셋인 LJ Speech 내 음성샘플 10개(16kHz, 10초 내외)를 활용했고, Google Colab 환경에서 PyTorch 및 torchaudio를 기반으로 구현하였다. STT 모델은 Wav2Vec2.0, TTS 모델은 XTTSv2 대상으로 제안기법 성능을 확인하였다.

평가지표는 다음 세 가지를 사용하였다.

- **WER(단어오류율):** 모델의 음성인식을 교란한 정도를 측정
- **Cosine Similarity(코사인 유사도):** DTW (Dynamic Time Warping)를 통해 시계열 정렬 후 원본과 합성 오디오의 음향적 차이를 판별
- **STOI(명료도):** 사람의 청각 기준에서 음성의 명료도 판단

4.2 실험 결과 및 분석

첫째, 제안기법을 적용시 STT, TTS 모델의 성능 지표가 모두 악화하는 것을 확인하였다. 평균 WER이 원본 오디오의 경우 0.054, 합성 오디오의 경우 0.224으로 다소 증가하여 설계한 적대적 노이즈가 음성처리 과정에서 모델의 학습에 오류를 유발함을 실험적으로 입증했다.

둘째, 제안기법은 원본 음성의 품질을 크게 훼손하지 않아 청취자가 들었을때 큰 영향을 주지 않는다. STOI 지표는 합성 오디오에서 0.90 수준으로 유지하였고, 코사인유사도 역시 0.89 수준으로 사람이 들었을 때 적대적 노이즈가 삽입된 음성이 원본 음성의 청취 품질과 거의 차이가 없음을 보여, 적대적 노이즈는 사람의 청각에는 영향을 주지 않았음을 확인했다.

[표 1] 평가지표별 실험결과

구분 (평균)	원본 음성	합성 음성(+적대적 노이즈)
WER	0.054	0.224
코사인유사도 (원본 vs 합성)	-	0.89
STOI	0.98	0.90

* WER: 0 ~ 1의 값으로 1에 가까울수록 오류가 심함을 의미
** 코사인유사도: -1 ~ 1의 값으로 1에 가까울수록 두 벡터가 유사함을 의미
*** STOI: 0 ~ 1의 값으로 1에 가까울수록 명료하게 들리며, 0.85 이상이면 명확하게 단어구분이 가능함을 의미

5. 결론 및 향후 연구

본 연구는 강연, 발표, 인터뷰 등 화자의 음성이 공개적으로 노출되는 상황에서 실시간으로 적대적 노이즈를 합성하여 음성 데이터의 무단 수집과 딥보이스의 학습을 사전에 차단할 수 있는 방어기법을 제안한다. 초도실험 결과, STT모델의 CTC 정렬 과정을 교란하고 TTS 모델의 Mel 특징 추출을 혼란시켜 음성처리 과정을 전체적으로 방해하며, 이를 통해 공격자가 확보한 음성 데이터의 모델 학습효과를 억제할 수 있음을 보였다.

향후 연구 방향은 다음과 같다. 첫째, 실험 데이터셋의 표본을 늘려 실증적 실험을 진행할 예정이다. 둘째, 제안기법을 통해 실제 회의, 강연 등 실제 환경에서 적용 가능성을 검증할 계획이다.

참고문헌

[1] Arık et al, "Deep voice: Real-time neural text-to-speech." International conference on machine learning. PMLR, 2017.

[2] Zwicker, Eberhard, and Hugo Fastl. “Psychoacoustics: Facts and models.” Vol. 22. Springer Science & Business Media, 2013.

[3] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572. 2014.

[4] Tabassi, Elham, et al, "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations.", NIST AI 100-2e2025, 2025.

[5] 한승우 et al, "Mel-Spectrogram과 MFCC를 이용한 딥러닝 기반 딥보이스 탐지시스템 개발에 관한 연구." 전기학회논문지 P 72.3. 2023.

[6] Fei et al, "VocalCrypt: Novel Active Defense Against Deepfake Voice Based on Masking Effect." arXiv preprint arXiv:2502.10329. 2025.

[7] Zeng et al, "Speech recognition with amplitude and frequency modulations." Proceedings of the National Academy of Sciences 102.7. 2005.