강원도 지역문제 해결을 위한 Personalized GAN 기반 딥보이스 피싱 탐지 기술

주제 선정 배경

강원특별자치도가 지키는 안심 명절..보이스피싱 없는 추석

김광현 기자

최근 보이스피싱 피해액은 매년 증가해 ('23년) 143억 원, ('24년) 151억 원, ('25년 4월) 52억 원으로 연말까지 160억 원 이상의 피해가 발생할 것으로 예상되고 있으며,

전체 피해자의 47% 이상이 60대 이상 고령층에 집중되고 있어 실질적 대응 능력을 높일 수 있는 현장형 교육이 절실한 상황이다.

강원특별자치도_지자체별 독거노인 현황 공공데이터 포털

A	B	C	D	E
시군	2020	2021	2022	2023
춘천	10058	11104	12319	13415
원주	11041	12454	13766	15336
강릉	9956	10813	11669	12639
동해	4218	4615	4979	5439
태백	2606	2875	3091	3351
속초	3877	4313	4658	5078
삼척	4047	4335	4641	4999
홍천	3999	4391	4752	5275
횡성	2811	3044	3240	3675
영월	2893	3082	3277	3569
평창	2795	3035	3214	3551
정선	2464	2625	2840	3161
철원	2066	2206	2312	2587
화천	1230	1356	1460	1610
양구	955	1026	1108	1237
인제	1443	1571	1706	1895
고성	1883	2006	2078	2308
양양	1902	2077	2239	2476

노인을 대상으로 하는 보이스 피싱 문제가 심화 되고 있으며

실제로 전체 피해자의 47%이상이 60대 이상 고령층에 집중되고 있다.

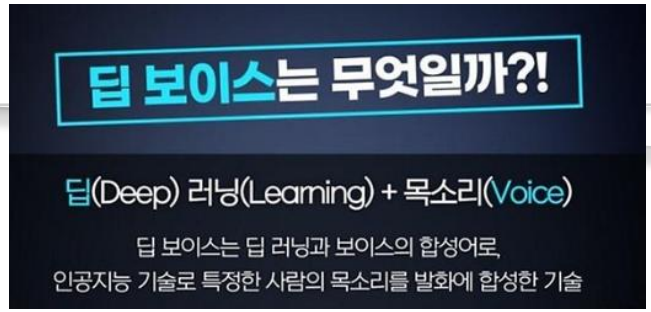
목적 및 필요성

딥보이스란?

딥보이스(Deep Voice)는 딥러닝을 이용해 특정인의 목소리를 똑같이 복제하는 AI 음성 합성 기술입니다.

단 몇 초의 실제 음성 샘플만 있어도, 그 사람의 억양, 감정, 말투까지 흉내 내어 새로운 문장을 말하게 할 수 있습니다.

이 때문에 실제와 구별이 거의 불가능하며, 최근 가족을 사칭하는 보이스피싱 범죄에 악용되어 심각한 사회 문제로 대두되고 있습니다.



70대 남성 임모씨는 납치당했다는 딸의 울먹이는 AI 딥보이스 전화에 속아 500만 원을 인출해 서울로 향함. 하지만 이는 목소리를 정교하게 위조한 '딥보이스 피싱' 사기였으며, 실제 딸은 전화한 적이 없는 것으로 밝혀짐.

최근 AI 딥보이스 피싱이 고령층의 감성과 신뢰를 공략하며 더욱 정교해지고 있으므로, 이들의 디지털 대응 한계를 고려하여 위협을 자동으로 탐지하고 즉각 차단하는 새로운 기술적 보호 장치 개발이 시급하다.

보이스 피싱 시나리오

보이스 피싱 대상 가족의 목소리 탈취 — 짧게는 5초에서 길게는 5분 이상의 발화 정보를 탈취

공격 모델 생성

실시간 음성 convert model 혹은 TTS 모델에 공격 대상자의 목소리

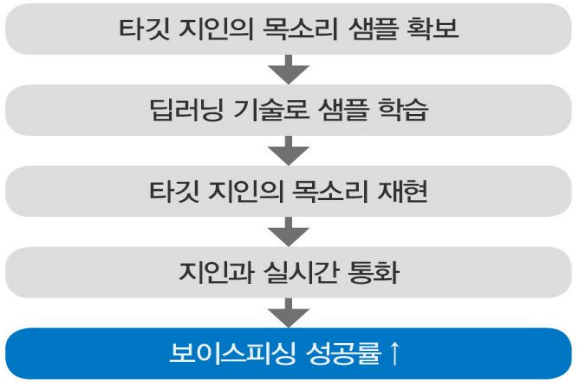
→ 공격자의 가족의 목소리로 피싱

예방 할 수 있는 방법

→ 공격모델 : 음성 변조 및 합성, 생성

→ 생성형 AI를 활용한 공격 모델 탐지 솔루션 개발

■ 딥보이스 보이스피싱 원리



■ 용어설명

☞딥보이스(Deep voice)
AI 학습 방식인 딥러닝(deep learning)과 목소리(voice)의 합성어. AI로 특정인의 목소리를 학습해 재현하는 기술.

통신망 – 음성 압축기술

MDCT

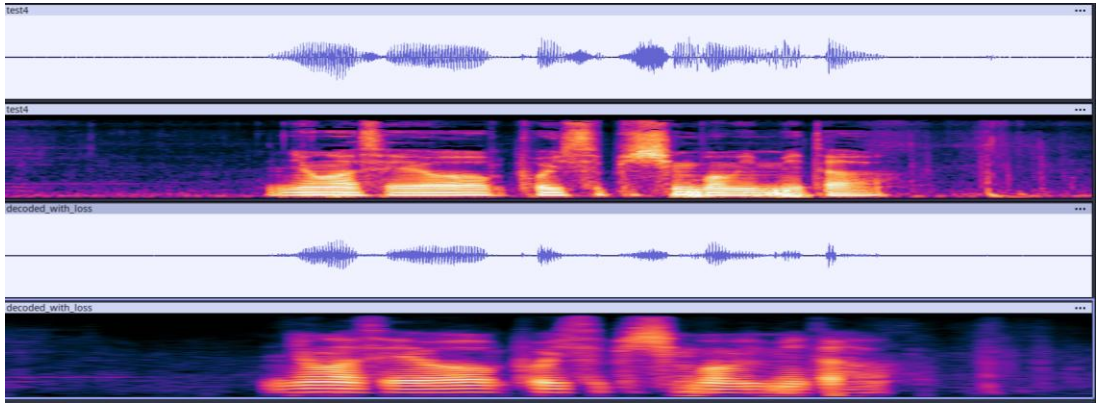
시간-주파수 도메인 변환 기법으로, 입력 신호를 50% 중첩된 윈도우로 분할하여 주파수 스펙트럼으로 변환
중첩 구간에서 TDAC특성을 활용해 역변환 시 완벽한 재구성이 가능하며, 블록 경계에서 발생하는 불연속성을 제거합니다.

Psychoacoustic Model

주파수 마스킹: 큰 소리 주변의 작은 소리는 들리지 않음
시간적 마스킹: 큰 소리 직전/직후의 작은 소리는 인지하기 어려움

Quantization

MDCT 계수를 스케일 팩터(scale factor)로 정규화한 후, 비선형 양자화 함수 (x^(3/4))를 적용하여 정수값으로 변환하는 과정입니다.



원본 음성과 통신망을 지나간 음성

실제 통신망의 패킷 손실과 압축을 보이기 위한 노이즈, 압축이 적용된 모습
OPUS 코덱을 사용, 강한신호 주변의 정보가 마스킹된 것과 패킷 손실되어 비어 있는 부분들을 확인 할 수 있다.

5초 단위로 음성파일을 나누고 학습 → 전처리와 데이터 증강 → 통신망을 거친 데이터셋 제작

Retrieval – Based Voice Conversion

Feature Extraction and Decoupling

입력 소스 음성에서 음높이(Pitch/F0), 리듬(Rhythm), 음색/발음(Timbre/Content)과 같은 핵심 특징들을 별도로 추출하고 분리

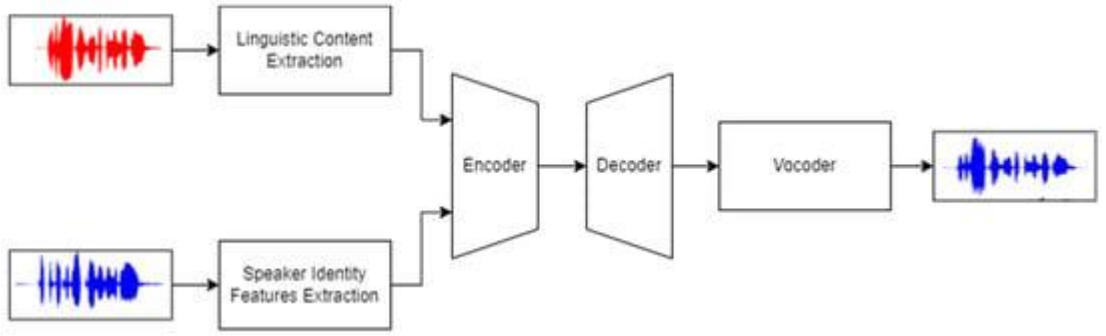
Feature Retrieval and Replacement

분리된 소스 음성의 음색/발음 특징 벡터를 대규모 목표 화자 데이터베이스에서 검색하여 가장 유사한 특징 벡터를 찾아 대체

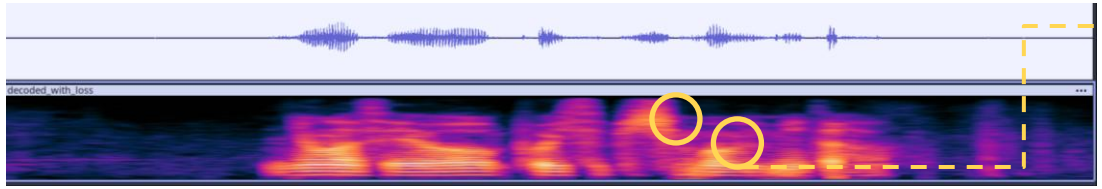
목표 화자의 음색을 가장 잘 나타내는 코드북또는 임베딩 벡터를 찾아내는 과정으로, 훈련 대신 검색(Retrieval)을 통해 음색 정보를 주입

Hybrid Synthesis

대체된 음색/발음 특징과 원본 소스 음성에서 가져온 음높이(Pitch) 및 리듬 정보를 결합



Retrieval – Based Voice Conversion



실시간 음성 변형에서 자주 보이는 노이즈 패턴들

우리는 이 패턴들을 학습하고 분류하는 모델을 개인 맞춤형으로 제공

Wav2Vec 2.0 + Linear Classifier

Conv Feature Encoder

원시 음성 신호(Raw Audio)를 입력받아 잠재적인 음성 특징 벡터를 추출하는 역할

인간의 청각 기관이 소리의 저수준 특징을 분리해내는 과정과 유사하며,
시간축을 따라 지역적 특징을 효과적으로 포착

Quantization

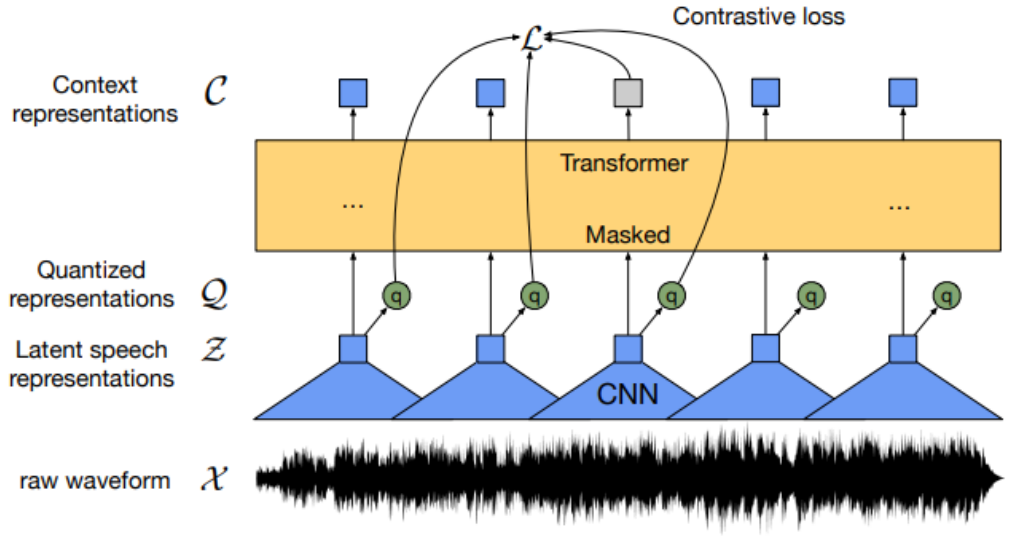
인코더가 추출한 연속적인 잠재 특징 벡터를 이산적인 음향 단위로 변환하는 단계

이산적인 단위를 사용하여 다음 핵심 개념인 대조 학습의 목표 레이블로 활용

Contrastive Learning Task

모델은 마스킹된 잠재 특징 벡터가 실제 올바른 양자화된 단위인지, 아니면 무작위로 추출된 다른 단위인지 구분하도록 훈련

자기 지도 학습의 핵심으로, 모델이 이산화된 목표 음향 단위를 예측하도록 학습하는 방식



wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, Michael Auli

wav2vec 2.0은 컨볼루션 인코더로 특징을 뽑고, 양자화 모듈로 이산적 목표를 만들며, 대조 학습을 통해 맥락을 이해하는 자기 지도 학습 프레임워크

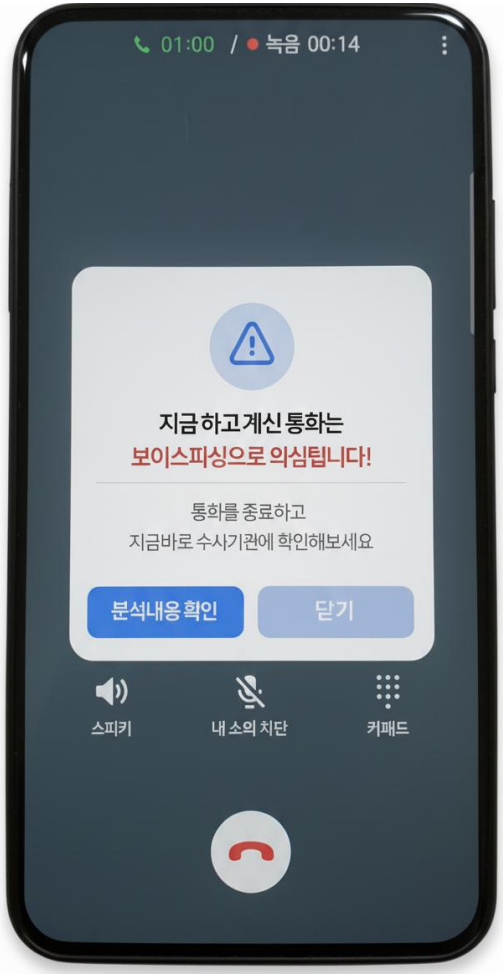
Wav2Vec 2.0

음성 데이터를 대조 학습하는 프레임 워크 생성된 음성의 노이즈 학습에 탁월한 성능

Linear Classifier

데이터 포인트를 분류하는 결정 경계를 직선 또는 평면으로 정의하는 모델

프로토 타입



학습 데이터

Retrieval-Based Voice Conversion 기반 생성된 FAKE 데이터

사용자 가족 혹은 지인의 실제 발화 데이터

Train Loss: 0.7001 | Train Acc: 48.34%
Test Loss: 0.6935 | Test Acc: 49.74% | AUC: 0.488

Classifier

상대방 음성 기반 분석 -> 분류

3초 ,5초 단위의 세그멘트로 분할하여 분류 진행

ELSE

통화 진행

이상치 탐지 결과 이상이 없으므로 딥보이스가 아니라고 판단하고 통하를 진행

IF : Classifier의 값이 임계 값 이상이면

FAST Api

온 디바이스 기반 종료 코드 생성과 전송

가벼운 API 제작이 가능한 FAST Api사용