Capstone Design Project

FISHER: Fraudulent Incoming Speech Handling and Event Recoder

Team Name: 20애기들

Name: 성영준, 오현택, 윤종우, 조민혁

Department: Mobile System Engineering

Student Number: 32202231, 32202733, 32202978, 32204292

목 차

<표 차례>

<그림 차례>

1. 프로젝트 개요

본 프로젝트는 실시간 통화 내용을 분석하여 보이스피싱을 탐지하고 사용자 편의를 위한 통화 요약 및 일정 자동 등록 기능, 사후 분석 레포트 생성 서비스 구축을 목표로 한다. 통화 중 발생하는 잠재적 위험 요소를 신속히 탐지하고 중요한 대화 정보를 추출·정리하여 향후 법정에서 활용될 수 있도록 타임라인 재구성을 통해 안전성과 효율성을 동시에 확보하고자 한다.

1-1. 제안 배경 및 필요성

최근 보이스피싱 범죄가 조직화·지능화되며 그 수법 또한 빠르게 고도화되고 있다. 특히 딥러닝 기반의 음성 합성 기술인 '딥보이스(Deep Voice)'의 발전으로 실제 사람의 목소리를 거의 완벽하게 묘사한 AI 합성 음성이 등장하면서 피해자가 가족이나 지인의 목소리를 쉽게 믿고 속는 사례가 증가하고 있다. 이러한 기술적 위협은 전화 통화 기반 금융사기, 개인정보 탈취, 심리적 협박 등 다양한 형태로 이어지고 있으며 기존의 수동 녹취나 사후 모니터링만으로는 효과적인 대응이 어렵다는 점에서 실시간 탐지 기술의 필요성이 절실하다.

한편, 일반 사용자들이 일상·업무 통화 중 놓치기 쉬운 중요한 정보(약속 일정, 요청사항 등)에 대한 관리 수요도 증가하고 있다. 현재 대부분의 사용자는 수동 메모나 통화 녹음을 통해 필요한 내용을 관리하고 있으나 이는 번거롭고 실효성이 낮은 경우가 많다. 이에 따라 통화 내용을 자동으로 텍스트화하고 주요 정보를 정리·요약해주는 기능에 대한 요구가 확대되고 있으며 특히 일정 관련 정보를 자동으로 캘린더에 등록하고 공유하는 기능은 일정 누락 방지와 사용자 편의성 제고에 크게 기여할 수 있다.

또한, 보이스피싱 발생 시 해당 통화 내용을 디지털 증거로 활용할 수 있는 포렌식 기능 역시 중요성이 커지고 있다. 본 프로젝트는 통화 중 수집된 음성 및 텍스트 데이터를 기반으로 위험 대화 발생 시점과 맥락을 자동으로 분석하고 통화 데이터에 대하여 해시 값을 생성함과 동시에 시간 순으로 재구성된 타임라인 레포트를 생성함으로써 통화 내용의 무결성을 확보하고 사후 증거자료로 법적 효력을 가질 수 있도록 한다. 이를 통해 단순 탐지·경고에 그치지 않고 실제 수사·재판 과정에서 활용 가능한 수준의 정량적·정성적 포렌식 정보를 제공함으로써 기존 탐지 기술과의 실질적 차별화를 구현하고자 한다.

따라서 본 프로젝트는 고도화되는 보이스피싱 위협에 대한 실시간 대응체계를 마련함과 동시에 통화 기반 정보의 체계적 관리, 포렌식 연계, 사용자 경험 향상을 위한 통합 솔루션 개발의 필요성에 기반하고 있다.

1-2. 프로젝트 목표

본 프로젝트의 주요 목표는 인공지능 기반 음성 분석 기술을 적용하여 실시간 통화 중 보이스피싱 여부를 탐지하고 통화 내용을 자동으로 기록·요약하며 일정 정보를 추출해 캘린더에 자동 등록·공유하는 통합형 모바일 애플리케이션 서비스를 개발하는 데 있다. 이를 통해 사용자에게 통화

보안과 정보 관리 편의성을 동시에 제공하고 기존의 수동적 대응을 넘어선 실시간 예방 중심의 스마트 음성 통화 환경을 구현하고자 한다.

첫째, 보이스피싱 예방을 위한 실시간 탐지 기능을 구현한다. 딥러닝 기반의 딥보이스 감지 모델과 위험 대화 탐지 알고리즘을 통합 적용하여 통화 중 금전 요구, 개인정보 요청, 긴급 상황 유도 등 의심스러운 발화 패턴을 자동으로 식별하고 사용자에게 실시간 경고 메시지를 제공함으로써 사기 피해를 사전에 차단할 수 있도록 한다. 또한 필요 시, 자동 번호 차단, 통화 중단, 위험 통화 자동 분류 등의 후속 조치 기능을 함께 제공하여 대응력을 강화한다.

둘째, 통화 정보에 대한 체계적인 관리 기능을 제공한다. 음성 인식(STT) 기술을 기반으로 통화 내용을 실시간 텍스트로 변환하고 대화의 주요 키워드를 중심으로 내용을 자동 요약함으로써 사용자가 통화 후에도 주요 논의사항을 쉽게 확인하고 기록할 수 있도록 한다. 이를 통해 업무 효율성을 높이고 통화 내 중요한 정보의 누락을 방지할 수 있다.

셋째, 통화 중 발생하는 일정 정보를 자동으로 추출하고 캘린더 연동 기능과 연계하여 등록 및 공유를 지원한다. 인식된 일정의 일시, 장소 등의 정보를 바탕으로 캘린더 일정이 자동 생성되며 관련 상대방에게는 카카오톡 또는 문자 메시지를 통해 일정 초대 및 리마인더 알림이 자동 발송되도록 구성하여 일정 관리의 실효성과 사용자 편의성을 높인다.

넷째, 보이스피싱이 실제로 발생했을 경우를 대비하여 디지털 포렌식 관점에서 통화 데이터를 증거로 활용할 수 있도록 한다. 통화 중 생성된 음성 및 텍스트 로그는 암호화 및 해시 처리를 통해 무결성을 확보하며, 이후 통화 흐름과 주요 발화 이벤트를 타임스탬프 기반 타임라인으로 자동 재구성함으로써 향후 수사기관이나 법정에서 디지털 증거로 활용 가능하도록 구조화된 리포트를 생성한다. 이는 단순 탐지 기능을 넘어, 정량적·정성적 분석이 가능한 포렌식 자료로써의 활용 가치를 지니며 기존의 탐지 도구들과는 명확히 구분되는 차별화된 기술적 기여점이라 할 수 있다.

궁극적으로 본 프로젝트는 AI 기반 실시간 보안 탐지 기술, 사용자 중심의 통화 편의 기능 그리고 법적 활용이 가능한 포렌식 기능을 결합한 통합형 서비스로서 보이스피싱 범죄 예방, 통화 정보 관리 자동화, 일정 연동, 디지털 증거화까지 아우르는 다기능 플랫폼을 구축하고자 한다. 이를 통해 사용자 개인의 보안 수준과 디지털 통화 경험을 동시에 향상시키는 것을 궁극적 목표로 한다.

2. 현황 및 문제점 분석

2-1. 보이스피싱 범죄 동향 및 사례 분석

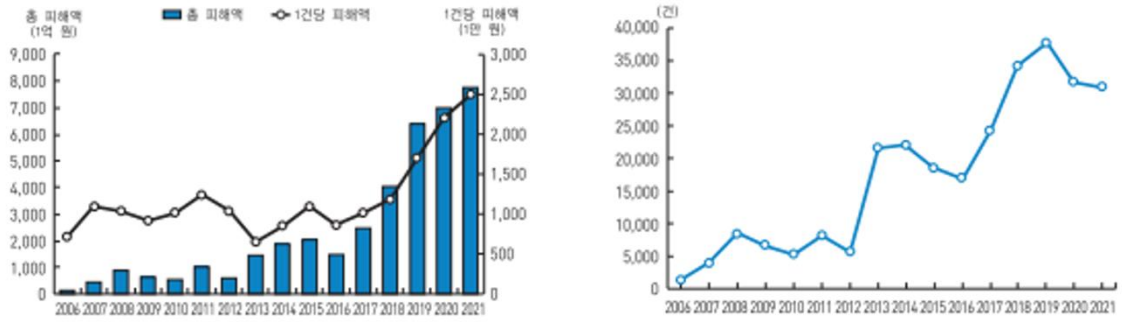

그림 1. 보이스피싱 피해금액 및 발생건수(2006-2021)

보이스피싱은 전화(Voice)와 개인정보(Phishing)를 합친 용어로 주로 전화나 문자메시지를 통해 공공기관, 금융기관, 가족 등을 사칭하여 개인 정보를 탈취하거나 금전적 피해를 입히는 사기 범죄로 그 수법이 날로 교묘해지고 있다.

먼저 보이스피싱의 주요 유형에는 두가지 기관 사칭형과 대출 사칭형이 있다. 각각을 살펴보면 먼저 기관 사칭형은 검찰, 경찰, 금융감독원 등의 공공기관을 사칭하여 피해자에게 범죄 연루 등을 이유로 금전을 요구하는 수법이다. 예를 들어 최근에는 악성코드를 활용하여 피해자의 전화를 가로채고, 가짜 공문서를 보내 신뢰를 얻은 후 금전을 편취하는 사례가 보고되다.

다음으로 대출 사기형인데 저금리 대출을 미끼로 피해자를 유인한 후 기존 대출 상환을 위해 현금 상환을 요구하거나 악성 앱 설치를 유도하여 금융 정보를 탈취하는 방식이다. 이처럼 다양한 형태로 우리 일상 속에 자연스럽게 침투하기 때문에, 자칫하면 자신도 모르는 사이에 피해를 입게 되며 그 피해 규모도 갈수록 커지고 있는 실정이다.

이에 따라 보이스피싱의 범죄 현황을 보다 구체적으로 살펴볼 필요가 있다. 그림 1과 같이 2006년 한 해에 1,488건의 보이스피싱 범죄가 발생하였다. 이 후 꾸준히 증가하여 2013년 2만 1,634건이 발생해 최초로2만 건을 돌파하였고, 2019년 3만 7,667건으로 한 해 발생 건수로는 최대를 기록하였다. 2006년부터 2021년까지 누적 발생 건수는 총 27만 8,200건이다. 보이스피싱으로 인한 피해 금액 또한 매우 심각한 수준인데 연도별 피해액을 살펴보면 2006년에는 106억원, 2018년 4,040억원, 2021년 7,744억원이 발생해 피해액이 계속 증가하고 있다. 이처럼 매년 사건 발생 건수는 다소 감소하거나 3만여 건에서 정체되어 있지만 피해액수가 계속 늘어난다는 점이 심각한 문제다. 즉 1건당 피해금액은 1,000만원 내외를 유지하다가 2019년 1,699만 원, 2020년 2,210만 원, 2021년에는 2,500만 원으로 최고를 기록하였다. 이는 발생 건수의 경우 관계기관의 노력 등으로 감소하였으나 범인들이 범행수법을 진화시키고 악성앱을 통해 피해자의 휴대폰을 원격 조종하게 되면서 피해 금액이 늘어난 것으로 분석할 수 있다. 2006년부터 2021년까지 누적 피해금액은 3조 8,681억원이다.

이처럼 피해가 해마다 심각해지는 상황에서 단순히 범죄 발생 이후의 사후 조치만으로는 보이스피싱 문제를 근본적으로 해결할 수 없다. 실제 범죄 예방을 위한 가장 효과적인 방법은 범죄 일당을 근원적으로 검거하는 것이다. 그러나 현실에서는 보이스피싱 범죄 조직의 상선 검거 비율이 고작 2%에 불과하다. 보이스피싱은 대부분 조직적으로 이루어지는 범죄이기 때문에 하부 조직원이 잡히더라도 꼬리 자르기 수법을 통해 상선은 도주하는 구조다.

2-2. 딥보이스 기술의 개념 및 악용 사례

딥보이스(Deep Voice) 기술은 딥러닝 기반의 고도화된 음성 합성 기술이다. 인공지능이 특정 인물의 음성 데이터를 학습하여 해당 인물의 말투, 억양, 호흡, 감정까지 모사한 새로운 음성을 생성할 수 있으며, 이는 일종의 음성 딥페이크 기술이라 할 수 있다. 기존의 TTS(Text-to-Speech) 기술이 단순히 문장을 기계적으로 읽는 수준에 머물렀다면, 딥러닝 기술의 도입 이후 음성 합성의 자연스러움과 현실감은 획기적으로 향상되었다.

최근에는 단 몇 초의 짧은 음성 샘플만으로도 화자의 목소리를 정밀하게 재현할 수 있을 정도로 기술이 고도화되었다. 이로 인해 발음, 억양, 감정 표현은 물론 말 사이의 침묵이나 숨소리 같은 비언어적 특성까지도 모사할 수 있게 되었으며 사람의 실제 육성처럼 들리는 수준의 합성이 가능해졌다.

이러한 기술은 다양한 분야에서 긍정적으로 활용되고 있다. 예를 들어, 엔터테인먼트 산업에서는 특정 음색을 복제하여 성별이나 연령이 다른 목소리 또는 유명인의 음성을 합성하는 데 활용되고 있으며 다국어 음성 콘텐츠 제작에도 응용되고 있다. 의료 및 복지 분야에서는 루게릭병과 같은 질환으로 음성을 잃은 환자가 생전의 본인 목소리를 학습시킨 AI 를 통해 다시 자신의 음성으로 소통할 수 있도록 지원하고 있다. 또한 시각장애인을 위한 음성 기반 독서 서비스, 맞춤형 음성 비서 등에서도 딥보이스 기술이 활용되며 접근성과 사용자 경험을 향상시키고 있다. 스마트폰 음성비서(Siri, Bixby 등)나 AI 스피커의 자연스러운 음성 안내도 딥보이스 기술 발전의 성과 중 하나이다.

그러나 이와 같은 기술의 발전은 동시에 심각한 악용 가능성도 내포하고 있다. 가장 대표적인 사례는 딥보이스 기술을 활용한 보이스피싱 범죄이다. AI 가 생성한 합성 음성은 실제 사람의 목소리와 구분이 어려울 정도로 정교하여 범죄자가 타인의 목소리를 복제해 지인이나 가족을 사칭하는 방식의 금융사기에 악용되고 있다. 특히 피해자가 감정적으로 동요하기 쉬운 상황을 노려 신속한 송금을 유도하는 수법이 자주 사용되고 있다.

실제로 2021 년 아랍에미리트(UAE)에서는 한 은행 담당자가 대기업 임원의 목소리로 걸려온 전화를 받고 약 420 억 원을 이체하였으나 이후 해당 음성이 AI 합성임이 밝혀졌다. 2023 년 캐나다에서는 부모가 아들의 다급한 목소리로 걸려온 전화에 속아 약 2 천만 원 상당의 비트코인을 송금한 사례가 발생했으며 이는 아들의 목소리를 흉내 낸 AI 음성을 활용한 보이스피싱으로 드러났다. 한국에서도 같은 해 고등학생 동생의 음성을 사칭한 딥보이스 피싱 사례가 보고되었으며 피해자는 6 천만 원을 송금한 뒤 뒤늦게 사기임을 인지하였다.

딥보이스를 활용한 이러한 범죄는 음성이라는 매체에 대한 신뢰 자체를 훼손시키며 전화 통화만으로는 더 이상 신원을 확인할 수 없다는 인식을 불러오고 있다. 경찰 당국 역시 모르는 번호로부터 걸려온 전화에 함부로 응답하지 말고 짧은 인사 한마디조차도 주의해야 한다고 경고하고 있다. 이는 짧은 음성 샘플조차 딥보이스 모델 학습에 이용될 수 있기 때문이다.

이처럼 딥보이스 기술은 새로운 형태의 보이스피싱 범죄 수단으로 부상하고 있으며 이에 대한 기술적·사회적 대응이 시급한 상황이다. 사용자 또한 전화 음성만으로 상대방을 신뢰하기보다 의심스러운 요청이 있을 경우 반드시 별도의 확인 절차를 거쳐야 하며 이에 대한 대중적인 인식 제고가 필요하다.

3. 프로젝트 범위 및 내용

3-1. 주요 기능 및 결과 화면 프로토타입

그림 2. 보이스피싱 및 딥보이스 탐지 기능 결과 화면

(1) 보이스피싱 및 딥보이스 탐지 및 경고

그림 2 는 보이스피싱 또는 딥보이스 음성이 식별되었을 때 사용자에게 경고 메시지를 제공하는 예시 화면을 보여준다. 이는 현재 프로토타입 화면이며 실제 구현 시에는 위험도를 확률값으로 산출하여 '정상', '의심', '확신' 3 단계로 구분해 사용자에게 안내할 계획이다. 예컨대 통화 중 딥보이스가 검출되거나 금전 요구·개인정보 요구와 같은 의심 발화 패턴이 발견될 경우 앱은 실시간으로 경고 알림을 띄워 사용자에게 즉각 주의를 환기시킨다. 이를 통해 보이스피싱 피해를 사전에 차단하고, 사용자가 신속히 대응할 수 있도록 지원한다.

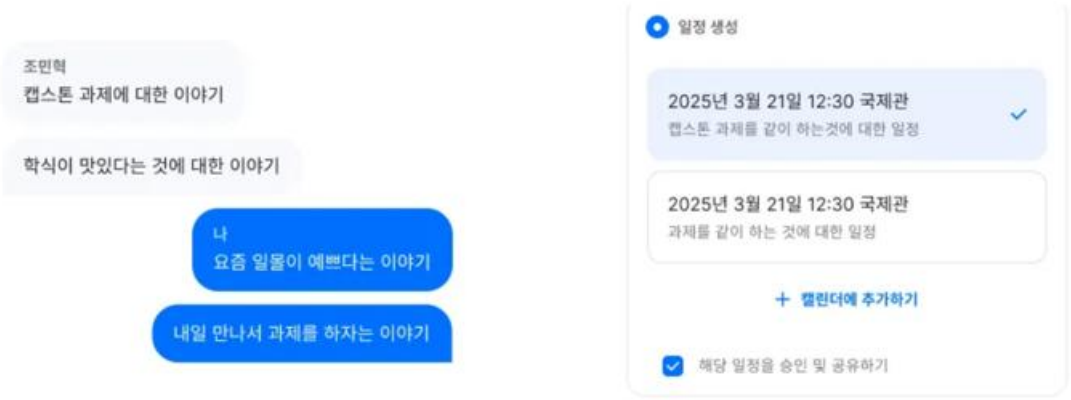

그림 3. 통화 내용을 추출/요약 및 일정 등록 기능 결과 화면

(2) 통화 내용 자동 요약 및 일정 등록

그림 3 은 통화 내용을 음성 인식(STT) 기술을 활용해 자동으로 텍스트로 변환하고 핵심 정보(일정·요청사항 등)를 추출하여 요약한 뒤 필요시 캘린더에 등록하는 과정을 시각화한 프로토타입 예시다. 단순 범죄 탐지 기능을 넘어 사용자가 통화 중 언급된 일정을 놓치지 않고 곧바로 등록·공유할 수 있도록 함으로써 통화 편의성을 높인다. 예컨대 약속 날짜와 장소, 참석자 정보를 인식해 자동으로 캘린더 이벤트를 생성하고 상대방에게 카카오톡 또는 문자 메시지로 일정 초대·리마인더를 발송할 수 있게 된다. 이를 통해 사용자 경험을 극대화하고 통화 후 별도의 메모나 앱 전환 없이 한 번에 일정 관리를 완료할 수 있다.

그림 4. 위험 번호 신고 기능 결과 화면

(3) 위험 번호 신고 기능

그림 4는 위험 번호 신고 기능의 프로토타입 결과 화면을 보여준다. 앱에서 특정 번호와의 통화에서 반복적으로 보이스피싱 의심 대화가 탐지될 경우, 해당 번호를 '위험 번호'로 분류·신고할 수 있게 하여 빠른 피해 방지와 사후 조치를 지원한다. 신고된 번호 정보는 내부 DB 혹은 스팸 전화 공유 플랫폼(예: 더콜)과 연동되어, 불특정 다수 사용자에게도 경고 알림이 제공될 수 있다. 이를 통해 보이스피싱 범죄 확산을 방지하고, 사용자의 안전성을 한층 높이는 효과를 기대할 수 있다.

Device Timestamp	Message	Estimated Time Value
2025-02-11 09:26:00 -> 2025-02-18 14:08:12	SystemClockTime: Setting time of day to sec=1739233560453	2025-02-18 14:08:12
2025-02-11 09:26:00 -> 2025-02-18 14:08:12	Auto time setting enabled: false	2025-02-18 14:08:12
2025-02-11 09:26:00 -> 2025-02-18 14:08:12	Before System Time : 2025-02-11 14:08:12	2025-02-18 14:08:12
2025-02-11 09:27:56 -> 2025-02-18 14:10:08	SMS Received to/from: 01049232198 Message: Hello AntiForesic	-

그림 5. 타임라인 재구성 레포트 결과 화면

(4) 타임라인 재구성 레포트

그림 5 는 통화 데이터를 기반으로 타임라인 형태로 재구성된 레포트의 예시 화면을 보여주며 이 역시 프로토타입 단계이다. 실제 구현 시에는 Timestamp 를 기준으로 통화 중 발생한 주요 발화(금전 요구 시점, 개인 정보 요청 발화 등)를 시간 순으로 재배치하여, 통화 내용 흐름을 직관적으로 확인할 수 있게 할 예정이다. 아울러 데이터 무결성 검증(암호화·해시 처리 등)을 통해 사후 법적 대응 시에도 증거 자료로 활용 가능하도록 설계한다. 이는 단순히 탐지 알림을 제공하는 것을 넘어, 보이스피싱 피해가 발생했을 때 해당 사실을 입증할 수 있는 구조화된 포렌식 리포트를 생성함으로써 프로젝트만의 차별화된 가치를 제공하는 기능이라 할 수 있다.

위와 같은 프로토타입 화면과 핵심 기능은 향후 구현 단계에서 UI/UX 개선, AI 탐지 정확도 향상, 법적 증거 활용성 제고 등을 고려하여 최적화될 계획이다. 이를 통해 본 프로젝트가 제시하는 통합 애플리케이션은 '보이스피싱 실시간 탐지 → 위험 여부 경고 → 통화 정보 요약 및 일정 관리 → 위험 번호 신고 및 포렌식 레포트'에 이르는 통화 솔루션을 제공함으로써 사용자 편의와 안전을 동시에 실현하는 것을 목표로 한다.

3-2. 시스템 아키텍처 프로토타입

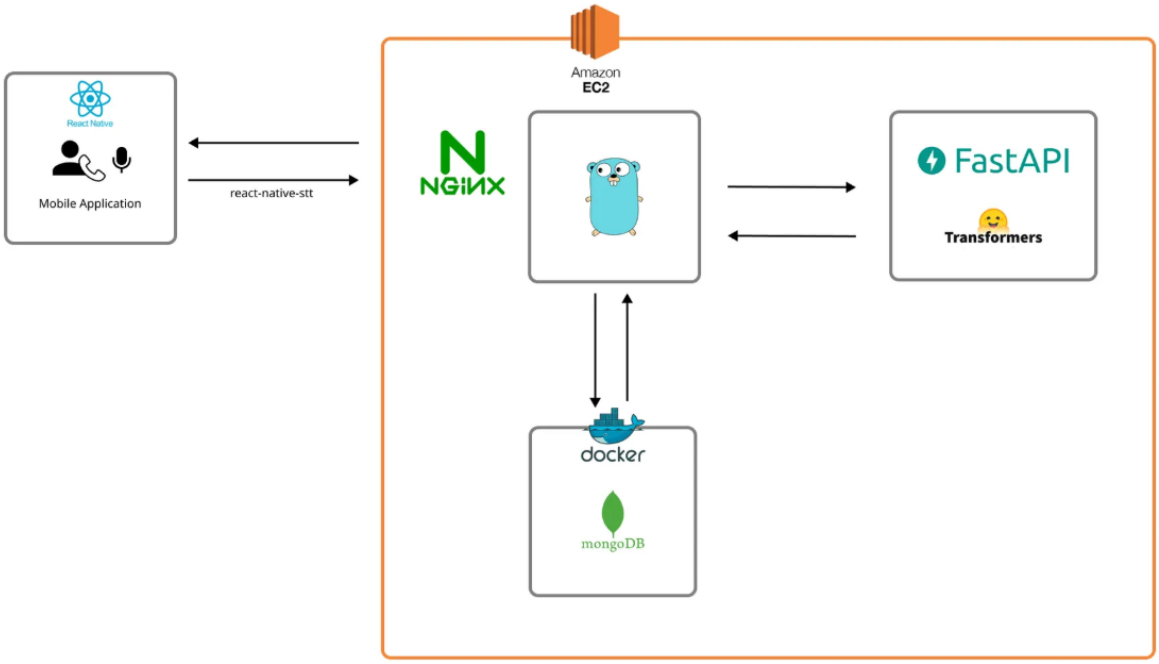

그림 6. 시스템 아키텍처 프로토타입

그림 6 은 본 프로젝트에서 제공하는 모바일 애플리케이션과 서버 측 구성 요소가 어떻게 연동되는지를 나타낸다. 전체적인 호스팅 환경은 Amazon EC2 인스턴스를 기반으로 운영되며 주요 기술 스택으로 React Native, Go, Nginx, FastAPI, Transformers, Docker, mongoDB 등을 활용한다. 사용자는 디바이스에 설치된 React Native 기반 애플리케이션을 통해 통화를 진행하고 react-native-stt 모듈을 통해 음성 데이터를 실시간으로 텍스트로 변환하거나 서버로 전송한다. 사용자 인터페이스는 보이스피싱 경고, 통화 요약, 일정 등록 등 다양한 기능을 시각화해 제공하며 프론트엔드에서 발생하는 모든 요청은 Amazon EC2 인스턴스 내에 위치한 Nginx 서버를 통해 처리된다. Nginx 는 요청의 라우팅을 담당하며 Go 기반 애플리케이션 또는 FastAPI 서버로 분기하여 각각의 로직에 맞는 처리를 수행하게 한다. 이 과정에서 대규모 트래픽 발생 시에는 로드 밸런싱 역할도 수행하여 서버의 안정성과 확장성을 보장한다. Go 서버는 본 프로젝트의 주요 비즈니스 로직을 담당하며 사용자 계정 관리, 통화 로그 저장, 보이스피싱 탐지 요청 처리 등을 수행하며 필요 시 FastAPI 서버와 연동하여 Transformers 와 같은 자연어 처리 엔진을 호출함으로써 복잡한 텍스트 분석, 딥보이스 판별, 요약 기능 등을 처리한다. FastAPI 는 Python 기반의 경량 웹 프레임워크로 Transformers 와 함께 동작하여 보이스피싱 의심 발화 탐지, 딥보이스 여부 판단, 중요 키워드 추출 및 일정 관련 정보 분류 등 고도화된 언어 분석 기능을 수행하며 처리 결과는 다시 Go 서버로 전달되어 사용자에게 실시간으로 알림이나 후속 조치를 제공한다. 서버 애플리케이션과 데이터베이스는 Docker 컨테이너로 패키징되어 운영되며 이를 통해 개발 환경과 배포 환경 간의 일관성을 유지하고, 필요 시 마이크로서비스 단위의 확장 및 배포가 용이하도록 구성된다. 데이터베이스는 mongoDB 기반으로 구성되며, 통화 로그, 사용자 정보, 일정 정보, 위험 번호 신고 내역 등 다양한 데이터를 안전하게 저장하고 관리하며 Docker 환경에서 구동됨으로써 이식성과 유지보수 편의성을 높인다. 모든 서버 자원은 Amazon EC2 상에서 운영되며

필요 시 오토 스케일링, 로드 모니터링, 로그 수집 기능을 함께 활용하여 안정적인 서비스 운영을 보장한다. 위와 같이 그림 6 의 시스템 아키텍처는 프론트엔드(Device) – 백엔드(Go, FastAPI) – 데이터베이스(mongoDB)가 유기적으로 연동되어 동작하며 음성 인식 기반 실시간 위험 탐지 기능과 사용자 중심의 통화 정보 관리 기능을 통합 제공함으로써 본 프로젝트가 지향하는 보이스피싱 예방, 일정 자동화, 법적 증거 활용까지 가능한 지능형 음성 통화 서비스의 기반을 구성한다.

3-3. 데이터 수집 및 학습 방법

3-3-1. 보이스피싱 데이터 수집 및 전처리

본 프로젝트에서 사용하는 보이스피싱 데이터는 금융감독원에서 제공하는 '보이스피싱 체험관'의 음성 파일과 유튜브에 공개된 실제 보이스피싱 피해 사례를 기반으로 수집된다. 이러한 데이터는 대부분 라벨이 부착되지 않은 상태이며 모델 학습을 위해 연구자가 직접 라벨링을 수행해야 한다. 이외에도 정상 음성 데이터는 AI Hub에서 제공하는 금융 상담 및 일반 한국어 대화 음성을 활용하며 보이스피싱 여부를 판단하는 모델 학습의 기반으로 사용된다. 수집된 모든 데이터는 Whisper 기반 자동 음성 인식 모델을 통해 텍스트화되고 이후 전처리 및 라벨링 단계를 거쳐 학습용으로 정제된다.

보이스피싱 탐지 모델은 오디오, 텍스트, 화자 정보 등을 함께 받아 각 시점(세그먼트)마다 보이스피싱 확률을 출력하는 구조이기 때문에 세그먼트 단위의 정밀한 라벨링이 필수적이다. 만약 통화 전체에 대해 단일 라벨(사기/정상)만 존재한다면 통화 전반에 대한 이진 분류 모델은 구축할 수 있으나 실시간 탐지 및 세부 위험도 분석 기능은 구현이 어렵다. 실제 보이스피싱은 통화 중반이나 후반부터 발생하는 경우가 많기 때문에 구간별 라벨링은 보다 정밀하고 실용적인 예측 성능을 확보하는 데 중요한 요소이다.

(1) 세그먼트 단위 설정 방식

라벨링을 위한 세그먼트는 일반적으로 두 가지 기준 중 하나로 나뉠 수 있다. 첫째는 3~5초 고정 길이 단위로 나누는 방식이다. 이 방식은 실시간 탐지 시스템과의 연동이 용이하고 구현이 간단하다는 장점이 있다. 그러나 문장 도중에 오디오가 분할되어 문장의 의미를 온전히 파악하지 못하거나, 발화자 교대가 일어나기 전에 세그먼트가 끊기면 화자 분리나 의미 분석에 오류가 발생할 수 있다.

둘째는 화자 교대(턴) 또는 질문, 답변 등 의미 단위 기준으로 나누는 방식이다. 이 방법은 각 세그먼트가 하나의 완결된 의미를 가지기 때문에 보다 정확한 라벨링이 가능하다는 장점이 있다. 다만, 세그먼트 길이가 일정하지 않아 실시간 시뮬레이션 시 처리 방식이 복잡해질 수 있으며, 때로는 한 턴이 너무 짧거나 길어지는 문제가 생긴다.

결론적으로, 본 프로젝트에서는 실시간 처리를 위한 고정 시간 단위 세그먼트 방식과 문장/화자 기준 분할을 절충하여 적용한다. 구체적으로는 3초 간격으로 Whisper를 실행하되, Whisper가 반환하는 문장 단위 또는 화자 턴 정보가 감지되면 해당 범위를 하나의 세그먼트로 간주하여 라벨링하는 혼합 방식이 가장 적절하다.

(2) 키워드 기반 반자동 라벨링 전략

모든 데이터를 수작업으로 라벨링하는 것은 많은 시간과 비용이 소요된다. 이를 보완하기 위해 금융 관련 키워드(‘계좌’, ‘송금’, ‘대출’)나 개인정보 관련 용어(‘주민번호’, ‘보안카드’, ‘공인인증서’ 등), 또는 수사기관 관련 키워드(‘조사’, ‘검찰’, ‘고소장’ 등)를 중심으로 자동 태깅 스크립트를 작성할 수 있다. 해당 키워드가 포함된 구간은 우선적으로 '의심 구간'으로 표시되고 이후 사람 라벨러가 직접 검토하여 최종 라벨을 부여한다. 이러한 방식은 라벨링의 효율성을 높이면서도 정확도를 일정 수준 이상으로 유지할 수 있는 현실적인 접근법이다.

구체적인 전처리 및 라벨링 프로세스는 다음과 같다:

1. Whisper로 전체 오디오를 텍스트화
2. 키워드 탐지 알고리즘을 통해 의심 구간 후보 도출
3. 의심 구간에 대해 사람이 세부 점검
4. 나머지 구간은 정상으로 간주하거나 샘플링하여 확인

이와 같은 반자동 라벨링 전략을 활용하면 전체 라벨링에 소요되는 시간과 리소스를 줄이면서도 실시간 구간 예측이 가능한 정밀한 학습 데이터를 구성할 수 있다.

3-3-2. 딥보이스 데이터 수집 및 전처리

본 프로젝트에서는 딥보이스를 탐지하기 위한 학습 데이터를 확보하기 위해 Fraunhofer CST, Zenodo, York University 등에서 제공하는 공개 음성 데이터셋을 수집한다. 이들 데이터셋에는 영어와 한국어를 포함해 다양한 언어 및 억양의 합성 음성이 포함되어 있으며 평균 250 시간 이상 분량의 음성 및 약 10 만 개 이상의 발화를 제공한다. 합성 음성 데이터는 인간 음성에 가까운 자연스러운 톤과 억양을 포괄하고 있어 모델이 보이스피싱에서 활용될 수 있는 실제적인 딥보이스 음성을 학습하는 데 큰 도움을 준다.

수집된 원시 음성 데이터는 서로 다른 샘플링 레이트와 녹음 환경(배경 소음, 마이크 품질 등)을 갖고 있어 모델 학습에 일관되게 활용하기 위해 전처리 과정을 수행한다. 첫째, 오디오 파일의 Sampling Rate 를 프로젝트 표준으로 통일하고, 고주파 대역을 필터링하여 통일된 스펙을 갖도록 정규화한다. 둘째, 배경 소음 제거를 위해 스펙트럼 감쇠나 밴드패스 필터 등의 기법을 적용해 과도한 잡음이 제거된 상태를 만든다. 셋째, 볼륨 정규화를 통해 음성 레벨 간 편차를 최소화하여 모델이 특정 발화의 크기에 치우치지 않고 안정적으로 특성을 학습할 수 있도록 한다.

이후 음성 신호에 대한 핵심 특징을 추출하기 위해 FFT(Fast Fourier Transform), Mel-Spectrogram, MFCC(Mel-Frequency Cepstral Coefficients) 등 다양한 신호 변환 과정을 수행한다. FFT 분석을 통해 음성의 주파수 분포를 확인하고, Mel-Spectrogram 은 인간이 인지하는 음역대를 반영한 스펙트럼 변환을 제공하며, MFCC 는 음성의 핵심 주파수 특성을 수치화하여 딥러닝 모델이 처리하기 적합한 입력으로 만든다. 이러한 특징들은 딥보이스가 만들어내는 특정 음색·억양 패턴을 모델이 효과적으로 구분해낼 수 있도록 도와준다.

한편, 수집된 데이터셋에는 실제 음성과 합성 음성이 섞여 있으므로, 이를 정확히 구분하여 모델에 라벨링 하는 작업이 필수적이다. 본 프로젝트에서는 각 음성 파일에 대해 ‘실제 음성’ 또는 ‘합성

음성’ 레이블을 명시적으로 부여하고, 필요 시 화자 정보(성별, 연령대, 억양 특성 등)와 같은 메타데이터도 함께 기록한다. 이렇게 구축된 라벨링 정보를 통해 모델은 어떤 음성이 합성되었는지 여부를 학습할 수 있으며, 추후 예측 과정에서 딥보이스 여부를 정확히 판별할 수 있게 된다.

전처리를 거친 최종 데이터셋은 훈련(train), 검증(validation), 테스트(test) 세 가지 세트로 분할하여 관리한다. 각 세트에는 다양한 환경·장치·언어가 혼합되도록 구성해 모델이 일반화 능력을 갖추도록 설계한다. 이와 같은 전처리 및 라벨링 과정을 통해 확보된 일관성 있는 음성 데이터는 딥보이스 탐지 모델의 학습·검증·평가 단계에서 안정적인 성능을 확보하는 기반이 된다.

4. 기술적 접근 방식

4-1. 보이스피싱 탐지 기법

보이스피싱을 효과적으로 탐지하기 위해, 본 시스템은 음성 기반의 데이터를 실시간으로 처리하고 분류하는 M-Module과 S-Module로 구성된 이중 파이프라인 구조를 따른다. 이 파이프라인은 오디오 입력을 받아 텍스트로 전환하고 화자 정보를 분리 및 정제한 뒤, 다중 모달 정보(오디오, 텍스트, 화자)를 통합하여 시계열 기반의 딥러닝 모델로 보이스피싱 여부를 판단하는 흐름을 따른다.

먼저 M-Module은 Whisper Small 모델을 이용하여 음성 데이터를 텍스트로 변환하는 역할을 수행한다. Whisper는 잡음 환경에서도 비교적 안정적인 인식 성능을 보이며, 짧은 음성 조각에서도 유의미한 결과를 도출할 수 있다는 점에서 본 프로젝트에 적합하다. 입력되는 음성은 3~5초 단위의 세그먼트이며, 향후에는 보이스피싱 특화 용어에 대한 인식을 강화하기 위한 파인튜닝도 고려될 수 있다.

전사된 텍스트는 이후 실시간 화자 분리(Speaker Diarization) 단계를 거치며 x-vector나 pyannote.audio와 같은 음향 임베딩 기반 알고리즘을 사용하여 발화자 간의 경계를 구분한다. 보이스피싱 통화는 일반적으로 '공격자'와 '피해자'라는 명확한 역할 구분이 존재하기 때문에 화자 분리는 이후 텍스트 분석에서 해당 화자의 말인지 식별할 수 있다는 점에서 중요한 역할을 한다.

화자 정보가 부착된 텍스트는 OpenAI GPT 기반의 Prompt Engineering을 통해 문장부호, 맞춤법, 구어체 보정 등 정제 작업을 거친다. 예를 들어 “어... 저기요... 그 이자금이…”와 같은 불완전한 발화를 “저기요, 이 자금이...”처럼 정돈된 형태로 변환할 수 있다. 이 과정에서 핵심 키워드(예: 계좌번호, 송금, 개인정보 등)를 자동 인식하거나 필요 시 해당 구간을 마스킹 및 태깅하여 이후 모델 학습에 도움이 되는 형태로 변환한다. 정제된 데이터는 각 발화 구간의 화자, 텍스트, 시간 정보와 함께 B-Module로 전달된다.

B-Module은 본격적인 보이스피싱 분류를 수행하는 모듈로 세 가지 유형의 임베딩을 활용한다. 첫 번째로 SEW(Split-Enhanced Wav2Vec2) 모델을 통해 오디오 임베딩을 추출한다. SEW는 기존 Wav2Vec2 대비 잡음에 강하며 짧은 음성 조각에서도 안정적으로 특징을 추출할 수 있다는 점에서 채택되었다. 두번째는 M-Module에서 얻은 정제된 텍스트는 KB-ALBERT 모델을 이용해 텍스트 임베딩으로 변환한다. KB-ALBERT는 한국어에 특화된 경량 BERT 계열 모델로 도메인 적합성과 추론

속도 측면에서 장점을 가진다. 세번째는 화자 정보를 별도의 텍스트 입력으로 임베딩하거나 유니크 토큰화를 통해 화자 임베딩을 생성한다.

이러한 임베딩 벡터들은 하나로 통합되어 Bi-LSTM 모델에 입력된다. Bi-LSTM은 시계열 정보를 고려한 구조로 과거 시점의 대화 흐름을 기억하면서 현재 구간이 보이스피싱일 가능성을 예측한다. 각 구간에 대해 보이스피싱 여부(0/1) 확률을 출력하며 학습 시에는 Cross Entropy 손실함수를 통해 라벨과의 오차를 최소화하도록 최적화된다. 예측된 확률이 특정 임계값(예: 70%)을 초과할 경우 보이스피싱 의심 경고를 발생시키도록 설정할 수 있으며 해당 임계값은 ROC 분석 등을 통해 조정 가능하다.

본 시스템은 짧은 세그먼트 단위의 실시간 탐지를 지원하면서도 화자별 발화 맥락과 음성적 특징을 통합적으로 고려할 수 있어 정확하고 실용적인 보이스피싱 탐지를 가능하게 한다.

4-2. 딥보이스 탐지 기법

본 프로젝트에서는 전처리가 완료된 음성 신호(Mel-Spectrogram, MFCC 등)를 입력으로 받아 합성 음성인지 실제 음성인지를 분류하는 이진 분류(binary classification) 모델을 구현하고자 한다. 이를 위해 대표적인 딥러닝 기반 접근 방식인 CNN, RNN, Transformer 세 가지 모델 구조를 비교·분석하며 각 기법은 음성의 시간적·주파수적 특징을 서로 다른 방식으로 학습함으로써 높은 정확도의 딥보이스 탐지를 가능하게 한다. 최종적으로는 훈련 단계에서 확보된 모델을 실시간 추론 환경에 적용하여 통화나 음성 녹취 중 발생할 수 있는 합성 음성을 빠르게 판별하고 경고를 제공하는 것을 목표로 한다.

1) CNN 기반 탐지 기법

CNN(Convolutional Neural Network)은 주로 이미지 처리에 사용되는 구조지만 음성 데이터를 2 차원 스펙트럼 이미지(Mel-Spectrogram, MFCC 등) 형태로 변환함으로써 효과적으로 적용할 수 있다. CNN 은 시간축과 주파수축을 기반으로 구성된 입력에서 합성곱 필터를 통해 로컬 패턴을 추출하며 이는 딥보이스가 만들어내는 비자연스러운 음색, 억양 변화, 고주파 왜곡 등의 특징을 포착하는 데 유리하다. 병렬 처리에 최적화되어 있어 학습 속도가 빠르고 대규모 데이터셋 처리에 효율적이며 VGGNet, ResNet 등 다양한 구조로 확장 가능한 유연성을 가진다. 다만 CNN 은 순차적 맥락이나 장기적인 시간 의존성을 직접적으로 반영하기에는 제한적이기 때문에 필요 시 RNN 또는 Attention 메커니즘과의 결합을 통해 시계열 정보를 보완할 수 있다. 실시간 추론 시에는 일정 길이로 나눈 음성 프레임별 스펙트럼을 CNN 에 입력하고 각 프레임의 합성 확률을 종합해 최종 판단을 내리는 슬라이딩 윈도우 방식을 적용할 수 있다.

2) RNN 기반 탐지 기법

RNN(Recurrent Neural Network)은 시계열 데이터 분석에 최적화된 구조로 시간에 따라 연속적으로 변하는 음성의 맥락을 학습하는 데 효과적이다. 특히 LSTM(Long Short-Term Memory)나 GRU(Gated Recurrent Unit) 같은 구조는 장기 의존성을 안정적으로 학습할 수 있어 음성의 흐름 속에서 발생하는 억양 전이, 비연속적인 합성 음성 구간, 감정 변화 등 딥보이스 특유의 패턴을 탐지하는 데 적합하다. RNN 은 CNN 에 비해 학습 속도가 느리고 병렬화에 제한이

있지만 Bidirectional RNN 을 적용하면 발화의 앞뒤 문맥을 동시에 반영할 수 있어 탐지 정확도를 크게 향상시킬 수 있다. 또한 Attention 메커니즘을 결합할 경우 긴 발화에서도 정보 손실 없이 중요한 부분을 강조할 수 있어 모델의 안정성과 성능을 강화할 수 있다. 실시간 환경에서 RNN 기반 모델을 사용하기 위해서는 음성을 일정 시간 단위로 끊어서 순차적으로 입력받고 디코딩 과정에서 각 시점의 합성 음성 여부 확률을 계산해 임계치를 초과하면 경고를 발동하는 구조를 구현할 수 있다.

3) Transformer 기반 탐지 기법

Transformer 는 Self-Attention 메커니즘을 기반으로 한 구조로 시퀀스 내의 모든 위치 간 상호작용을 동시에 학습할 수 있어 긴 시간 범위에 걸친 의존성도 효율적으로 모델링할 수 있다. 최근 음성 분야에서는 Transformer 를 기반으로 한 다양한 구조(Speech-Transformer, Conformer, wav2vec 2.0 등)가 등장하며 음성 인식과 합성 탐지 분야에서 높은 성능을 보이고 있다. 본 프로젝트에서는 Mel-Spectrogram 이나 MFCC 와 같은 전처리된 음성 데이터를 시퀀스 형태로 Transformer 인코더에 입력하여 발화 전체에 걸쳐 나타나는 합성 음성의 특징을 포착한다. 실시간 처리를 위해서는 일정 길이의 음성 시퀀스를 누적한 뒤 모델에 입력하여 다음 구간으로 넘어가기 전 합성 음성 확률을 추론하는 방법을 사용할 수 있다. 긴 문맥까지 고려해야 할 경우 Conformer 등의 개선된 모델로 발화 흐름 전체를 효율적으로 캡처하는 방식을 채택할 수 있다.

이와 같이 CNN, RNN, Transformer 는 각각의 구조적 특성과 강점을 바탕으로 딥보이스 탐지에 효과적으로 활용될 수 있으며 본 프로젝트에서는 세 가지 접근 방식 모두를 훈련 단계에서 실험적으로 적용해 정확도, 탐지 속도, 연산 효율성을 비교한 후 최적의 모델을 선정하고자 한다. 실시간 추론 단계에서는 전처리된 음성 프레임 또는 시퀀스를 모델에 입력해 합성 확률을 산출한 뒤 특정 임계치를 초과하면 사용자에게 경고 알림을 띄우고 통화 로그나 포렌식 기록에 해당 사건을 저장하는 방식으로 동작한다. 또한 필요 시 CNN 과 RNN 또는 Transformer 를 결합한 하이브리드 모델을 적용함으로써 시간적 문맥 이해와 스펙트럼 기반 특징 추출을 동시에 강화하고 실시간 환경에서도 높은 성능을 발휘할 수 있는 딥보이스 탐지 시스템을 구축할 계획이다. 이러한 최적 모델을 서비스화하여 상시 동작하는 보이스피싱 방어 체계를 마련하고 추후 포렌식 보고서 생성, 법정 증거 활용까지 연계되는 확장적인 보안 솔루션을 구현할 계획이다.

4-3. 수집 데이터 포렌식 기법 및 활용

본 프로젝트에서는 보이스피싱 범죄 대응 및 사후 분석(포렌식) 기능을 강화하기 위해, 통화 과정에서 생성된 데이터로부터 타임라인 재구성 레포트를 도출한다. 이를 위해 먼저, 사용자 기기에서 연락처에 없는 번호나 스팸 의심 번호로부터 전화가 걸려올 경우 자동으로 로그 생성을 시작한다. 통화 진행 중에는 특정 시간 간격으로 타임스탬프가 포함된 로그 데이터를 텍스트 파일 형식으로 기기 로컬에 누적 기록하며, 통화 종료 후에는 해당 로그 파일에 대한 해시값을 생성해 위·변조 여부를 확인할 수 있도록 한다.

생성된 텍스트 파일과 해시값은 보안 채널을 통해 서버로 전송되며, 서버 측에서는 전송받은 텍스트 파일에 대해 해시값을 생성하여 전송받은 해시값과 일치하는 지 확인하여 일치하면 변조되지 않았다

판단해 암호화를 통해 안전한 스토리지에 저장함으로써 무결성을 보장한다. 기기 로컬에는 불필요한 파일이 장시간 남아있지 않도록 서버 전송이 완료된 뒤에는 로컬에 저장된 텍스트 파일을 삭제함으로써 사용자 디바이스의 내부 저장소를 효율적으로 관리한다. 이후, 특정 통화 이벤트가 의심 또는 확정 단계의 보이스피싱 사례로 판단될 경우 서버에 저장된 해시값과 로그 파일의 해시값이 일치하는지 검증하여 무결성을 한 번 더 확인한다.

무결성이 확인된 로그 파일은 타임라인 재구성 알고리즘을 통해 발화 시점, 대화 내용, 의심 발화 구간 등을 시계열로 재정렬하여 타임라인 레포트로 생성된다. 이 레포트에는 주요 이벤트(금전 요구, 계좌번호 요청 등)의 발생 시간과 대화 맥락이 명확히 표시되어 수사기관이나 법정에서 디지털 증거로 활용될 수 있도록 한다. 특히 해시 기반의 체인 오브 커스터디(Chain of Custody)를 구축해 로그 파일이 생성된 시점부터 법적 활용 시점까지의 변조 여부를 투명하게 증명하는 점이 본 프로젝트의 포렌식 기능에서 중요한 차별화 요소이다.

결과적으로, 이러한 포렌식 작업 흐름은 실시간 탐지와 함께 사후 증거 확보에 이르는 완결형 보안 솔루션을 지향하며 향후 보이스피싱 피해에 대한 신속한 법적 대응과 범죄 예방 측면에서 상당한 파급효과를 기대할 수 있다.

5. 프로젝트 관리 및 성능 평가

5-1. 일정 및 리소스 계획

표 1. 일정 및 리소스 계획

단계	기간 (주차 기준)	주요 작업 내용
1 단계 요구사항 분석 및 아키텍처 설계(약 2 주)	**4 월 1 주 ~ 4 월 2 주**(4/1 ~ 4/14)	- 딥보이스·LNN 모델 요구사항 정의 - 통화 녹음/STT 프로세스 설계 - 일정 자동 등록 기능(API) 구조 설계 - 보안 체계(암호화·개인정보 정책) 수립
2 단계 프로토타입 개발 및 초기 검증(약 3~4 주)	4 월 3 주 ~ 5 월 1 주(4/15 ~ 5/5)	- 보이스피싱 모니터링 프로토타입 구현(딥보이스 검출, 위험 키워드 탐지) -잭 STT 엔진 연동 및 기본 요약 기능 구현 - 캘린더 등록 프로세스(상용 API) 프로토타입 - 내부 테스트 시나리오 기반 정확도·속도 측정 및 우선 개선
3 단계 기능 고도화 및 통합 테스트(약 2 주)	5 월 2 주 ~ 5 월 3 주(5/6 ~ 5/19)	- 보이스피싱 대응 로직 고도화(위험 등급별 알림/차단 시나리오) - 통화 요약 알고리즘 개선(주요 아젠다·약속 인식·사용자 키워드 반영) - 약속 공유·자동 초대 기능 통합 -

		통합 테스트(성능·부하·보안 취약점 점검) 실시
4 단계 시범 운영 및 개선(약 1 주 이상)	5 월 4 주(5/20 ~ 5/26)	- 내부 임직원·베타테스터 대상 시범 운영 - 사용자 피드백 수집 후 위험 탐지 정확도·사용성·안정성 개선 - 개인정보 암호화 범위 확대, 접근 제어 정책 재점검
5 단계 배포 및 운영(지속)	5 월 4 주 이후	- 긴급 패치 및 추가 기능 업데이트 - 최신 피싱 트렌드 반영, 위험 키워드 지속 업데이트 - 사용자 맞춤형 설정 기능 고도화(관심 일정 유형 확대 등)

본 프로젝트는 AI 모델 개발과 시스템 구현 및 운영을 위해 전문적인 인력을 구성하여 역할을 명확히 분담한다. 조민혁과 윤종우는 AI 분야에서 딥보이스 모델과 LNN 기반의 모델 연구개발을 수행하여 음성 합성 탐지 정확성을 높이고 보이스피싱을 효과적으로 탐지할 수 있도록 한다. 오현택과 성영준은 백엔드 개발을 맡아 통화 녹음 처리, 실시간 음성 텍스트 변환(STT) 서비스 연동, API 설계를 통해 안정적인 서버 환경을 구축한다. 또한 성영준과 오현택은 프론트엔드 개발에도 참여하여 사용자 친화적인 UI/UX 설계와 알림, 일정 관리 기능을 구현해 사용자 편의성을 최대화한다.

실시간 음성처리를 위해서는 클라우드 기반 서버를 적극 활용하여 유연하고 확장성 있는 시스템을 구성할 계획이다. AI 모델의 학습과 추론을 위한 고성능 GPU 서버를 확보하여 실시간 탐지 성능을 보장하고 사용자 데이터 관리를 위한 DB 서버는 암호화와 접근 제어가 철저히 이루어진 보안성이 높은 환경으로 구축된다.

5-2. 위험 요인 및 대응 전략

AI 모델의 성능이 부족할 경우 실제 보이스피싱을 탐지하지 못하거나 정상 통화를 잘못 탐지하는 등의 문제가 발생할 수 있다. 이에 따라 다양한 환경의 음성 데이터 수집과 데이터셋 확장을 지속적으로 수행하고 최신 합성 음성 기술을 반영하여 주기적으로 모델을 재학습한다.

개인정보 유출 위험을 예방하기 위해 데이터 전송 단계부터 SSL/TLS 기반의 암호화 프로토콜을 적용하며 저장된 데이터 또한 강력한 암호화 기술을 통해 보호한다. 데이터 접근 권한을 엄격하게 제한하고, 지속적인 로그 모니터링과 정기적 보안 점검을 통해 보안 위험을 최소화한다.

실시간 처리 과정에서 발생할 수 있는 속도 저하 문제는 경량화된 모델과 분산 처리 기술을 도입하여 극복한다. 또한 스트리밍 기반의 설계를 통해 중요 구간을 우선 분석하여 실시간 탐지 효율을 높이고 고성능 GPU 서버를 활용하여 AI 추론 속도를 최적화한다.

마지막으로 사용자 오용 및 악용 사례를 방지하기 위해 서비스 이용 시 개인정보 처리방침과 이용약관을 통해 법적 책임을 명확히 안내한다. 특히 통화 녹음 시 상대방의 동의를 얻도록 정책을 마련하고 악용이 감지된 경우 신속히 계정 차단 및 법적 조치를 안내하여 사용자와 서비스 모두를 보호할

계획이다.

6. 기대효과 및 활용 방안

6-1. 일반 사용자 보호 및 대중 인식 제고

본 시스템은 AI 기반의 실시간 음성 분석을 통해 보이스피싱 범죄를 조기에 탐지하여 사용자에게 즉각적인 경고를 제공함으로써 금융 피해를 예방한다. 특히 최근 급증하는 딥보이스(Deepfake Voice) 기반의 보이스피싱에도 대응할 수 있도록 합성 음성 여부를 판별하는 기술을 적용하여 최신화된 범죄 수법에 효과적으로 대응할 수 있다. 이는 기존의 사후적 대응 방식에서 벗어나 능동적이고 선제적인 피해 예방 체계를 구축하는 데 기여할 것이다.

또한 음성 인식(STT) 기술을 활용해 통화 내용을 자동으로 텍스트로 변환하고 이를 통해 주요 내용을 빠르게 요약하여 사용자가 중요한 정보를 놓치지 않도록 지원한다. 통화 중 언급된 일정이나 약속 정보는 자동으로 캘린더와 연동하여 사용자 편의성을 크게 증대시키며 보다 체계적이고 효율적인 일상 및 업무 관리가 가능해진다.

더불어 개별 사용자의 통화 데이터를 분석하여 맞춤형 보안 가이드를 제공하고 보이스피싱 위험 패턴을 시각화하여 사용자가 스스로 통화 습관을 분석할 수 있도록 함으로써 지속적인 보안 인식 제고와 금융 범죄 예방의 효과를 극대화할 수 있다.

6-2. 금융/통신/공공기관 활용 방안

본 시스템은 금융기관과 통신사가 고객 보호를 위한 보이스피싱 예방 시스템으로 활용하여 고객 신뢰와 브랜드 이미지를 강화할 수 있다. 특히 고객 상담 시스템과 IVR 서비스 내에 실시간 탐지 기술을 도입하면 보이스피싱 의심 통화를 즉각적으로 대응할 수 있어 금융 피해 예방의 실효성을 높일 수 있다.

공공기관과 보안기업은 본 시스템에서 수집된 데이터를 바탕으로 보이스피싱 범죄 패턴을 분석하여 효과적인 예방 캠페인을 수립할 수 있다. 경찰청 및 금융감독원 등 관련 기관과의 긴밀한 협력을 통해 범죄 추적 및 사후 대응력을 높이고 사이버 보안 기업과의 기술 연계를 통해 보안 솔루션을 상용화하고 표준화하는 데에도 기여할 것이다.

콜센터 및 고객 상담 부서에서는 실시간 보이스피싱 탐지 결과를 상담사에게 제공하여 고객과의 통화 중 긴급 상황을 신속하게 대처할 수 있게 한다. 통화 내용을 자동 기록 및 요약하는 기능은 상담 품질을 높이고, 금융 범죄 발생 시 법적 대응에 필요한 증거자료로 유용하게 활용할 수 있다.

7. 결론 및 향후 발전 방향

7-1. 프로젝트 기대 성과

본 프로젝트는 보이스피싱에 대한 실시간 탐지를 가능하게 하는 AI 기반 시스템을 통해, 사전 예방 중심의 새로운 보안 패러다임을 제시하고자 한다. 특히 Whisper 기반의 음성 인식, 화자 분리 및 GPT 기반 텍스트 정제, 그리고 Bi-LSTM을 활용한 시계열 기반 분류 구조는 통화 중 발생할 수 있는 다양한 형태의 보이스피싱을 효과적으로 포착할 수 있도록 한다.

더불어, 딥보이스(Deepfake Voice) 탐지 기술을 통해 최근 고도화된 AI 음성 합성 기반 보이스피싱에도 대응 가능하다는 점에서 본 시스템은 기존 기술들보다 더욱 강력한 대응력을 확보하고 있다. 이러한 점에서 본 프로젝트는 단순한 기술적 구현을 넘어 실질적인 사회적 피해 감소와 사용자 인식 제고라는 두 가지 측면에서 중요한 성과를 기대할 수 있다.

7-2. 법적·윤리적 고려 사항

본 시스템은 통화 내용 및 음성 데이터를 기반으로 학습 및 추론을 수행하므로 개인정보 보호 및 통신 비밀 보장과 관련된 법적, 윤리적 고려가 필수적이다. 특히 사용자 음성의 실시간 처리 과정에서 발생할 수 있는 정보 수집, 저장, 활용 범위에 대해 명확한 고지 및 동의 절차가 필요하며 비식별화 및 암호화 등의 기술적 조치를 통해 개인정보 보호 수준을 강화해야 한다.

또한 딥보이스 탐지 기능이 포함됨에 따라 향후 음성 진위 여부와 관련된 법적 분쟁이 발생할 수 있음에 유의해야 하며, 탐지 결과에 대한 투명성과 해석 가능성을 확보하는 것이 중요하다. 나아가 보이스피싱 탐지 결과를 타인에게 자동 공유하거나 공공기관에 자동 보고하는 구조를 도입할 경우 반드시 해당 사용자에 대한 명확한 동의와 법적 근거가 수반되어야 한다.

이와 같은 윤리적·법적 검토를 바탕으로 본 시스템은 기술의 효용성과 사용자 권리 보호 간의 균형을 이룬 책임 있는 인공지능 서비스로 구현되어야 한다.

8. References

[1] 내 목소리 빼앗는 '딥보이스 피싱' 주의보,
http://www.boannews.com/media/view.asp?idx=132751

[2] "'형! 살려줘' 당신 목소리가 범죄에 쓰인다: AI 보이스피싱의 덫",
https://v.daum.net/v/20240421093843598

[3] 보이스피싱 범행단계별 대응방안 연구,
https://www.kisa.or.kr/20301/form?lang_type=KO&page=&postSeq=29#fnPostAttachDownload

[4] Milandu Keith Moussavou Boussougou, "머신러닝 기법을 이용한 한국어 보이스피싱 텍스트 분류 성능 분석", ACK 2021 학술발표대회 논문집 (28권 2호)

[5] Jiangyan Yi, et al. "Audio Deepfake Detection: A Survey", JOURNAL OF LATEX CLASS FILES, VOL. 14, NO. 8, AUGUST 2023

[6] Qian Luo, K.V Sivasundari, "Whisper+AASIST for DeepFake Audio Detection", HCII 2024

[7] P. Kawa, M. Plata, et al. "Improved DeepFake Detection Using Whisper Features", INTERSPEECH 2023

[8] J. J. Brid, Ahmad Lotfi, "REAL-TIME DETECTION OF AI-GENERATED SPEECH FOR DEEPFAKE VOICE CONVERSION", arXiv:2308.12734

[9] S. Reardon, "How Deepfake Voice Detection Works",
https://www.pindrop.com/article/deepfake-voice-detection/

[10] N. Q. Do, A. Selamat, et al. "Deep Learning for Phishing Detection: Taxonomy, Current Challenges and Future Directions", IEEE Access

[11] N.A Bhaskaran, M. Srinadh, et al. "Detecting Deep Fake Voice using Machine Learning", 10.34293/sijash.v11iS3-July.7919

"