고령자 보이스피싱 예방을 위한 인공지능 기반 음성합성 훈련 플랫폼 개발

정유태[1], 조광문[2*]
[1]국립목포대학교 융합소프트웨어학과 학생, [2]국립목포대학교 컴퓨터학부 교수

Development of an AI-Based Voice Synthesis Training Platform for Voice Phishing Prevention on Older Adults

Yu-Tae Jung[1], Kwangmoon Cho[2*]
[1]Student, Dept. of Software Convergence Engineering, Mokpo National University
[2]Professor, School of Computer Science and Engineering, Mokpo National University

요 약 보이스피싱 범죄는 고령자를 주요 대상으로 심각한 경제적·정신적 피해를 초래하고 있으며, 최근 인공지능 음성합성 기술의 발전으로 실제 인물의 음성을 모방한 범죄 수법이 고도화되면서 기존의 정보 제공 중심 예방 교육의 한계가 드러나고 있다. 본 연구는 고령자 보이스피싱 예방 훈련을 위해 인공지능 음성합성 학습모델과 음성 데이터 수집 플랫폼을 설계·구현하였다. 제안한 시스템은 고령자가 모바일 애플리케이션을 통해 직접 음성 데이터를 수집하고, 소량의 음성 데이터 환경에서도 작동 가능한 음성합성 모델을 활용하여 개인화된 음성 기반 훈련 콘텐츠를 생성하는 구조로 구성된다. 또한 훈련 플랫폼은 청취-판단-피드백의 학습 흐름을 기반으로 설계되어, 고령자가 반복적인 체험을 통해 보이스피싱 상황 인식 능력을 향상시킬 수 있도록 하였다. 시스템 구현 및 동작 검증 결과, 음성 데이터 수집부터 모델 학습, 음성 합성, 훈련 콘텐츠 제공에 이르는 전 과정이 안정적으로 수행됨을 확인하였다. 본 연구는 소량의 고령자 음성 데이터를 활용한 음성합성 기반 보이스피싱 예방 훈련 시스템의 기술적 구현 가능성을 검증하였으며, 향후 고령자 대상 체험형 예방 교육 시스템 개발을 위한 기초 자료로 활용될 수 있을 것으로 기대된다.

주제어 : 보이스피싱, 고령자, 음성합성, AI 학습모델, 소량 데이터 학습, 예방 훈련 시스템

Abstract Voice phishing has caused serious economic and psychological damage, particularly targeting older adults. With recent advances in AI-based voice synthesis technologies, voice phishing attacks that imitate real individuals have become increasingly sophisticated, revealing the limitations of conventional prevention education based primarily on information delivery. This study designs and implements an integrated platform for voice phishing prevention training for older adults using an AI-based voice synthesis learning model and a voice data collection system. The proposed system enables older adults to collect voice data directly through a mobile application and generates personalized voice-based training content using a voice synthesis model that operates effectively in low-resource data environments. In addition, the training platform is designed based on a three-stage learning flow of "listening-judgment-feedback," allowing older adults to improve their awareness of voice phishing scenarios through repeated experiential training. System implementation and functional verification confirm that the entire pipeline—from voice data collection and model training to voice synthesis and training content delivery—operates stably as intended. This study verifies the technical feasibility of a voice synthesis–based voice phishing prevention training system using a limited amount of older adult voice data and is expected to serve as foundational work for developing experiential prevention education systems for older adults in the future.

Key Words : Voice Phishing, Older Adult, Voice Synthesis, AI Learning Model, Prevention Training System

본 과제(결과물)는 2025년도 교육부 및 전라남도의 재원으로 전라남도RISE센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과입니다.(2025-RISE-14-001)
*교신저자 : 조광문(Kwangmoon.Cho@gmail.com)
접수일 2025년 12월 22일 수정일 2026년 01월 13일 심사완료일 2026년 01월 26일

1. 서론

보이스피싱 범죄는 정보통신 기술의 발달과 함께 지속적으로 진화하며, 사회 전반에 걸쳐 심각한 경제적·정신적 피해를 초래하고 있다[1]. 특히 전화 통화나 음성 메시지를 기반으로 한 범죄는 대면 상황과 유사한 신뢰 환경을 형성하여 피해자의 경계심을 낮추는 특징을 가지며, 이로 인해 사회적 약자를 집중적으로 노리는 구조적 범죄 형태로 인식되고 있다[2].

보이스피싱 피해자 중에서도 고령자는 인지 처리 속도의 저하, 디지털 기기 사용 경험 부족 등으로 인해 범죄 상황을 즉각적으로 판단하고 대응하는 데 어려움이 있는 취약 계층이다[3]. 특히 가족이나 공공기관 관계자를 사칭하는 음성 기반 접근 방식은 고령자의 심리적 신뢰를 빠르게 형성하여 피해로 이어질 가능성을 높인다[3].

최근 인공지능 음성합성 기술의 발전은 이러한 범죄 양상을 더욱 복잡하게 만들고 있다[1]. 딥러닝 기반 음성합성 기술은 소량의 음성 데이터만으로도 특정 화자의 음성을 자연스럽게 재현할 수 있으며, 이는 실제 인물의 음성을 모방한 이른바 딥보이스 기반 범죄로 악용될 수 있다[4]. 이러한 음성은 기존의 기계적인 사기 음성보다 자연스러워, 특히 고령자가 진위를 판단하는 데 큰 어려움을 겪을 수 있다[5].

기존의 보이스피싱 예방 교육은 문자 안내나 홍보 영상과 같은 정보 제공 중심의 방식에 머물러 있어, 실제 범죄 상황에서의 대응 능력을 향상시키는 데 한계가 있다[6]. 이에 따라 최근에는 실제 상황을 모사한 체험형·실습형 예방 교육의 필요성이 강조되고 있다[6].

한편 인공지능 음성합성 기술은 범죄에 악용될 수 있는 위험성을 지니는 동시에, 이를 적절히 활용할 경우 보이스피싱 상황을 안전하게 재현할 수 있는 예방 훈련 도구로 활용될 수 있다[1]. 그러나 기존 연구들은 주로 UI/UX 설계나 교육 콘텐츠 구성에 초점을 두고 있으며, 실제 고령자의 음성 데이터를 수집하여 이를 기반으로 음성합성 모델을 학습하고 훈련 시스템에 적용한 사례는 상대적으로 부족하다[6].

이에 본 논문에서는 고령자 보이스피싱 예방 훈련을 위한 인공지능 음성합성 학습모델을 설계하고, 이를 지원하는 음성 데이터 수집 플랫폼을 개발한다. 제안하는 시스템은 고령자가 직접 음성을 녹음하여 학습 데이터를 생성하고, 소량의 음성 데이터 환경에서도 적용 가능한 음성합성 모델을 통해 개인화된 음성 기반 훈련 환경을 제공한다. 본 연구의 주요 기여는 고령자 특성을 고려한 음성 데이터 수집 구조 제안, 소량 데이터 환경에 적합한 음성합성 학습 파이프라인 설계, 그리고 통합형 보이스피싱 예방 훈련 시스템의 구현 및 동작 검증에 있다.

2. 관련 연구

2.1 보이스피싱 예방 교육 연구

보이스피싱 예방을 목적으로 한 기존 연구들은 주로 범죄 유형 분석과 예방 수칙 제시에 초점을 맞추어 왔다. 이러한 연구들은 보이스피싱 범죄의 사회적 위험성을 알리고, 일반 사용자에게 경각심을 부여하는 데 일정 부분 기여하였다. 그러나 대부분의 연구가 정보 전달 중심의 접근 방식을 취하고 있어, 실제 범죄 상황에서의 대응 능력 향상이라는 측면에서는 한계를 가진다[7].

최근에는 이러한 한계를 보완하기 위해 상황 기반 시뮬레이션이나 역할극 형태의 예방 교육 방식이 제안되고 있다. 선행 연구에서는 실제 피싱 공격을 모사한 시뮬레이션 환경이나 역할극 기반 훈련이 사용자의 위험 인식과 대응 행동을 유의미하게 향상시킬 수 있음을 보고하였다[8]. 그러나 이러한 접근은 주로 이메일 기반 피싱이나 텍스트 중심 시나리오에 초점을 두고 있으며, 음성 기반 보이스피싱 상황을 충분히 반영하지 못한다는 한계를 가진다. 특히 음성 자극의 현실성이 부족할 경우 실제 범죄 상황과의 괴리가 발생하여, 실질적인 훈련 효과 확보에는 제약이 따른다.

2.2 고령자 친화형 인터페이스 연구

고령자 친화형 인터페이스에 대한 연구는 주로 가독성 확보, 조작 단순화, 인지 부담 감소를 중심으로 진행되어 왔다. 기존 연구에 따르면 글자 크기 확대, 명확한 색 대비, 화면 요소의 최소화는 고령자의 정보 인식 정확도를 향상시키는 데 긍정적인 영향을 미친다[9,10]. 또한 터치 영역을 충분히 확보하고 화면 전환 횟수를 줄이는 설계는 고령자의 조작 오류를 감소시키는 것으로 보고되고 있다[11].

이와 더불어 입력 방식 역시 고령자 서비스 설계에서 중요한 요소로 다루어진다. 자유 입력 방식은 입력 오류와 사용 부담을 증가시키는 반면, 선택형 입력 방식은 사용자의 인지 부담을 줄이고 데이터의 일관성을 확보하는 데 효과적인 것으로 알려져 있다[9-11]. 이러한 연구 결과들은 본 논문에서 제안하는 음성 데이터 수집 플랫폼의 설계 방향을 설정하는 데 중요한 근거로 활용되었다.

2.3 음성합성 기술 및 소량 데이터 학습 연구

음성합성(TTS) 기술은 딥러닝 기반 모델의 발전과 함께 자연스러운 음성 생성을 가능하게 하였으며, 다양한 서비스 분야에서 활용되고 있다. 그러나 대부분의 고품질 음성합성 모델은 화자별로 대량의 음성 데이터를 요구하며, 이는 고령자 음성 데이터를 대상으로 적용하기에는 현실적인 제약이 따른다[12].

이러한 한계를 극복하기 위해 최근에는 소량의 음성 데이터만으로도 화자 특성을 반영할 수 있는 few-shot 및 zero-shot 음성합성 모델이 제안되고 있다[13,14]. 이러한 모델들은 제한된 데이터 환경에서도 비교적 안정적인 음성 합성을 가능하게 하며, 개인화된 음성 서비스 구현에 적합한 특성을 가진다. 그러나 기존 연구들은 주로 기술적 성능 평가에 초점을 맞추고 있으며, 실제 교육 또는 훈련 시스템에 적용한 사례는 많지 않다[15].

2.4 기존 연구의 한계 및 본 연구의 위치

앞서 살펴본 바와 같이, 기존 보이스피싱 예방 교육 연구는 주로 이메일 또는 텍스트 기반 시뮬레이션에 집중되어 있으며, 음성 기반 보이스피싱 상황과 고령자 특성을 동시에 고려한 연구는 제한적이다. 특히 고령자의 실제 음성 데이터를 수집하고 이를 기반으로 AI 음성합성 모델을 학습하여 예방 훈련에 적용한 통합 시스템에 대한 연구는 부족한 실정이다[15].

본 연구는 이러한 연구 공백을 해소하고자, 고령자 음성 데이터 수집, 인공지능 음성합성 학습, 보이스피싱 예방 훈련을 하나의 플랫폼으로 통합하여 설계·구현한다는 점에서 기존 연구와 차별성을 가진다.

3. 시스템 전체 구조 및 요구사항 분석

본 연구에서 제안하는 보이스피싱 예방 훈련 시스템은 고령자의 사용 특성과 음성 데이터 학습 환경을 동시에 고려하여 설계되었다. 음성 데이터 수집부터 인공지능 모델 학습, 훈련 콘텐츠 제공까지의 전 과정을 하나의 흐름으로 연결하는 것을 목표로 하였으며, 이를 위해 사용자 요구사항과 시스템 요구사항을 종합적으로 분석하였다.

3.1 사용자 및 시스템 요구사항 분석

고령자를 주요 사용자로 하는 시스템은 조작 단계 단순화와 직관적인 인터페이스 제공이 핵심 요구사항으로 도출되었다[10]. 특히 음성 녹음과 같이 사용자가 직접 수행해야 하는 기능은 현재 상태를 명확히 인지할 수 있도록 시각적 피드백을 제공할 필요가 있다.

시스템 관점에서는 음성 데이터의 안정적인 저장 및 관리, 모델 학습 요청 처리, 그리고 소량 데이터 환경에서도 학습이 가능한 구조가 주요 요구사항으로 도출되었다.

3.2 시스템 구조 설계 대안

시스템 구조 설계 단계에서는 처리 위치에 따라 세 가지 대안을 검토하였다. 첫째, 모든 처리를 모바일 애플리케이션에서 수행하는 클라이언트 중심 구조, 둘째, 모든 데이터 처리와 모델 학습을 서버에서 수행하는 서버 중심 구조, 셋째, 모바일 애플리케이션과 서버가 역할을 분담하는 클라이언트-서버 분리 구조이다.

각 구조는 연산 효율성, 사용자 인터페이스 설계, 시스템 확장성 측면에서 서로 다른 장단점을 가진다.

3.3 최종 시스템 구조

본 연구에서는 앞서 검토한 구조 중 클라이언트-서버 분리 구조를 최종 시스템 구조로 채택하였다. 해당 구조에서 모바일 애플리케이션은 음성 데이터 수집과 사용자 인터페이스를 담당하며, 서버는 음성 데이터 처리, 인공지능 음성합성 모델 학습, 학습된 모델 관리 기능을 수행한다.

사용자가 앱을 통해 음성을 녹음하면 데이터는 서버로 전송되어 저장 및 전처리 과정을 거친 후 모델 학습에 활용되며, 학습 완료된 모델은 음성 합성 및 훈련 콘텐츠 제공에 사용된다. 이러한 구조를 통해 음성 데이터 수집-학습-활용 과정이 하나의 순환 흐름으로 작동하도록 설계하였다.

4. 고령자 음성 데이터 수집 및 관리 설계

음성합성 기반 보이스피싱 예방 훈련 시스템에서 음성 데이터는 시스템 성능과 활용 가능성을 결정하는 핵심 요소이다. 특히 고령자 음성은 일반 성인 음성과는 다른 발화 특성을 가지므로, 데이터 수집 단계에서부터 이에 대한 고려가 필요하다. 본 장에서는 고령자 음성 데이터의 특성과 기존 데이터셋의 한계를 분석하고, 이를 반영한 데이터 수집 및 관리 전략을 제시한다.

4.1 고령자 음성 데이터의 특수성

고령자의 음성은 발화 속도가 느리거나 불규칙하며, 억양과 발음의 일관성이 낮은 경우가 많다[16]. 또한 발음 정확도가 낮거나 쉼과 망설임이 빈번하게 나타나는 경향이 있어, 일반적인 음성합성 모델 학습 과정에서 잡음 요소로 작용할 가능성이 있다. 이러한 특성으로 인해 고령자 음성 데이터를 일반 성인 음성과 동일한 방식으로 처리할 경우 학습 성능 저하로 이어질 수 있다.

한편 기존 공개 음성 데이터셋은 대부분 방송 또는 스튜디오 환경에서 수집된 성인 화자 음성을 중심으로 구성되어 있어, 고령자의 실제 발화 특성을 충분히 반영하지 못한다. 또한 연령대 편중 문제와 개인정보 보호 및 활용 범위 제약으로 인해, 실제 교육·훈련 목적의 시스템에 직접 적용하는 데 한계가 존재한다. 이로 인해 고령자 음성 데이터를 대상으로 한 학습을 위해서는 훈련 목적에 부합하는 별도의 데이터 수집 전략이 요구된다.

4.2 본 연구의 음성 데이터 수집 전략

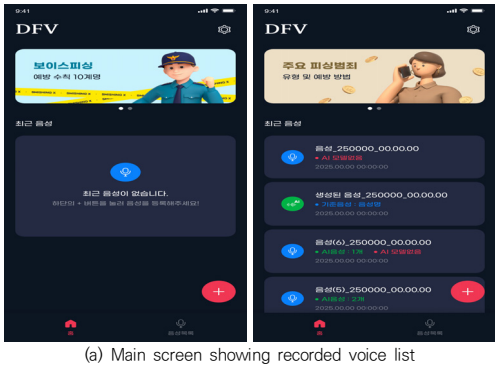
본 연구에서는 화자당 음성 길이를 최소화하되 화자 수를 확보하는 전략을 채택하였다. 이는 개별 화자의 음성 특성을 과도하게 학습하기보다는, 고령자 음성의 전반적인 특성을 반영하는 데 목적을 둔 것이다. 사용자는 모바일 애플리케이션을 통해 짧은 문장을 여러 차례 녹음하며, 녹음 과정은 고령자의 조작 부담을 최소화할 수 있도록 단순한 인터페이스로 설계되었다.

4.3 메타데이터 설계 및 데이터 관리

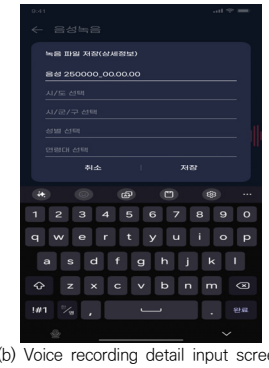
본 연구를 위해 구축된 고령자 음성 데이터 셋은 총 50명의 화자(60대 37명, 70대 10명, 80대 2명, 90대 1명)로부터 수집되었으며, 전체 규모는 약 250분 분량의 1,500 문장으로 구성된다.

수집된 음성 데이터는 파일 식별자와 함께 연령대, 성별, 지역 정보를 포함한 메타데이터로 관리된다. 모든 메타데이터는 선택형 입력 방식으로 수집하여 데이터의 일관성을 확보하였으며, 실명이나 연락처와 같은 민감한 개인정보는 수집하지 않았다. 모든 데이터는 익명화된 식별자를 통해 관리되며, 이는 향후 데이터 확장 및 추가 분석을 고려한 관리 구조를 제공한다.

제안하는 모바일 애플리케이션은 이러한 데이터 수집 구조를 바탕으로 메인 화면에서 보이스피싱 예방 배너와 최근 음성 리스트를 한 화면에 제공하고 사용자가 한눈에 재생 및 관리할 수 있도록 한다. 음성 녹음 상세 화면에서는 녹음 후 이름, 지역, 성별, 연령대 등의 기본 정보를 선택형 입력으로 받도록 설계하여 고령자가 과도한 텍스트 입력 없이도 메타데이터를 손쉽게 입력할 수 있도록 한다.



(a) Main screen showing recorded voice list



(b) Voice recording detail input screen

[Fig. 1] Voice data collection interfaces of the proposed mobile platform

5. 인공지능 음성합성 학습모델 설계 및 학습 전략

본 연구의 음성합성 모델 설계는 “소량의 고령자 음성 데이터 환경에서 실제 활용이 가능한가”라는 문제의식을 바탕으로 설계되었다. 일반적인 음성합성 연구에서는 고품질 음성 생성 자체가 주요 목표로 설정되는 경우가 많지만, 본 연구의 목적은 보이스피싱 예방 훈련이라는 응용 시나리오에 적합한 음성 생성이 가능한지를 검증하는 데 있다. 즉, 방송 수준의 음질이나 정밀한 화자 복제보다는, 실제 보이스피싱 상황을 충분히 모사할 수 있는 수준의 음성 자연성과 현실감을 제공하는 것이 더 중요하다고 판단하였다.

또한 본 연구는 고령자 음성 데이터라는 특수한 환경을 전제로 하고 있으며, 화자당 확보 가능한 음성 데이터의 양이 제한적이라는 현실적인 제약 조건을 가진다. 따라서 모델 설계 단계에서부터 대규모 데이터 기반의 음성합성 접근 방식보다는, 소량 데이터 환경에서도 안정적으로 동작할 수 있는 학습 구조를 중심으로 검토를 진행하였다.

5.1 소량 데이터 학습의 이론적 배경

기존의 TTS(Text-to-Speech) 모델은 화자 고유의 발화 특성과 음색을 안정적으로 학습하기 위해 비교적 많은 양의 음성 데이터를 필요로 한다[17]. 일반적으로 수십 분에서 수 시간에 이르는 화자별 음성 데이터가 요구되며, 이러한 조건은 고령자 음성 데이터 수집 환경에서는 현실적으로 충족하기 어렵다. 고령자의 경우 장시간 녹음에 대한 피로도가 높고, 반복적인 발화 수행이 부담으로 작용할 수 있기 때문이다.

이러한 한계를 극복하기 위한 대안으로 최근에는 few-shot 및 zero-shot 음성합성 학습 방식이 주목받고 있다[18]. few-shot 학습은 소량의 화자 음성을 기반으로 사전 학습된 모델을 미세 조정하여 화자 특성을 반영하는 방식이며, zero-shot 학습은 별도의 화자별 학습 없이도 참조 음성을 통해 화자 특성을 추론하는 방식이다. 이러한 접근 방식은 데이터 확보가 어려운 환경에서도 음성합성 모델을 적용할 수 있는 가능성을 제공한다.

본 연구에서는 이러한 소량 데이터 학습 개념이 고령자 음성 데이터 환경에 적합하다고 판단하였으며, 모델 선택 및 학습 전략 설계의 핵심 이론적 배경으로 활용하였다.

5.2 후보 모델 비교 분석

음성합성 모델 선정을 위해 본 연구에서는 대표적인 TTS 계열 모델들을 비교·분석하였다. 우선 Tacotron 계열 모델은 텍스트-음성 매핑 구조가 비교적 단순하며, 자연스러운 음성을 생성할 수 있다는 장점을 가진다. 그러나 화자 특성 학습을 위해서는 충분한 양의 음성 데이터가 필요하며, 소량 데이터 환경에서는 학습 안정성이 떨어질 가능성이 있다.

VITS 계열 모델은 end-to-end 구조를 기반으로 음질과 발화 자연성 측면에서 우수한 성능을 보인다. 그러나 이 역시 일정 수준 이상의 데이터가 확보되지 않을 경우, 화자 특성 반영이 불안정해질 수 있다는 한계를 가진다. 특히 고령자 음성과 같이 발화 변동성이 큰 데이터 환경에서는 학습 실패 가능성이 존재한다.

반면 XTTS 계열 모델은 대규모 음성 데이터로 사전 학습된 구조를 기반으로 하며, fine-tuning과 zero-shot 음성 생성을 동시에 지원한다는 특징을 가진다. 이는 소량의 화자 음성 데이터만으로도 화자 특성을 일정 수준 반영할 수 있음을 의미한다. 본 연구에서 확보한 화자당 약 5분 분량, 30문장 수준의 데이터는 일반적인 TTS 모델 학습에는 부족한 양이지만, XTTS 모델의 설계 특성상 적용 가능성이 있다고 판단하였다. 이러한 이유로 본 연구에서는 XTTS 계열 모델을 최종적으로 선택하였으며 이하에서는 XTTS v2.0.2의 학습구조와 본 연구에서의 적용 방법을 정리한다. 또한 소량 데이터 환경에서의 실제 적용 가능성은 이후 절에서 제시하는 학습 결과 및 정량 평가를 통해 검증하였다.

5.3 학습 전략 및 파라미터 설정

XTTS v2.0.2는 텍스트를 직접 음성 파형으로 변환하는 단일 단계 구조가 아닌 텍스트 입력으로부터 음향 토큰 시퀀스를 예측하는 단계와 예측된 토큰을 오디오 파형으로 복원하는 보코더 단계를 갖는 2단계 구조를 따른다. 입력 텍스트 시퀀스는 다음과 같이 정의한다.

$$X = (X_1, X_2, \cdots, X_T)$$

여기서 X_T는 텍스트 토큰이며 T는 텍스트 길이를 의미한다. 정답 음향 표현은 멜 스펙트로그램이 아닌 사전 학습된 오디오 코덱을 통해 얻은 이산 음향 토큰 시퀀스로 정의한다.

$$Z = (Z_1, Z_2, \cdots, Z_N)$$

여기서 z_t는 음향 토큰을 의미하며 N은 음향 토큰 길이이다. GPT 기반 자기회귀 모델은 다음과 같은 조건부 확률을 모델링한다.

$$P(Z|X) = \prod_{t=1}^{N} P(z_t | z_{<t}, X)$$

학습 목표는 정답 음향 토큰 시퀀스 Z와 모델이 예측한 확률 분포 간의 교차 엔트로피 손실을 최소화하는 것으로 정의된다.

$$L_{CE} = -\sum_{t=1}^{N} \log P(z_t | z_{<t}, X)$$

본 연구에서는 사전 학습된 XTTS v2.0.2를 초기값으로 사용하고 GPT 기반 음향 토큰 예측 모델에 대해서만 미세 조정을 수행하였으며 보코더는 사전 학습 가중치를

유지하였다. 예측된 음향 토큰 시퀀스를 $\hat{Z}$ 라 할 때 최종 음성 파형 $\hat{W}$는 다음과 같이 정의된다.

$$\hat{W} = Vocoder(\hat{Z})$$

이와 같은 구조는 모델이 대규모 다화자 데이터로부터 이미 학습한 일반적인 음성-텍스트 정렬 능력을 유지한 채 소량의 화자 데이터로 화자 특성 분포만을 미세 조정하도록 한다. 따라서 전체 모델을 처음부터 학습하는 방식에 비해 필요한 데이터 규모가 현저히 감소한다. 모델 학습은 Tesla V100-PCIE-32GB GPU 환경에서 수행되었으며 epoch 수를 20으로 제한하고 batch size는 8로 설정하였다. Optimizer는 AdamW를 사용하였고 초기 학습률은 5e-6으로 설정하였다. 그 외 학습 세부 설정은 AllTalk TTS의 기본 구성을 따랐으며 화자 1명당 평균 학습 소요 시간은 약 10~15분으로 측정되었다.

5.4 학습 결과의 활용 가능성 분석

학습된 음성합성 모델은 고령자 음성의 발화 속도, 억양, 음색 특성을 일정 수준 반영하는 음성 생성을 가능하게 하였다. 생성된 음성은 실제 보이스피싱 상황을 모사한 훈련 콘텐츠로 활용하기에 충분한 자연성을 제공하는 것으로 관찰되었다.

본 연구에서는 음질의 절대적인 우수성보다는, 훈련 대상자가 보이스피싱 음성으로 인식할 수 있는 현실감을 제공하는지를 주요 평가 기준으로 삼았다. 이에 따라 본 절에서는 정성적인 관점에서 훈련 컨텐츠로서의 활용가능성을 논의하고 정량적인 음성 명료도 평가는 7.4절에서 별도로 다룬다.

6. 보이스피싱 예방 훈련 플랫폼 구현

본 연구에서 구현한 보이스피싱 예방 훈련 플랫폼은 음성합성 결과 제공에 그치지 않고, 사용자가 보이스피싱 상황을 인식하고 판단하는 과정을 경험할 수 있도록 설계된 훈련 중심 구조를 가진다. 플랫폼은 실제 범죄 상황과 유사한 음성 자극을 제공함으로써, 고령자가 반복적인 훈련을 통해 보이스피싱 음성의 특징을 인지하도록 하는 것을 목표로 한다.

6.1 훈련 플랫폼의 교육적 설계 목적

본 연구에서는 보이스피싱 예방 교육의 효과를 높이기 위해 “청취–판단–피드백”의 3단계 학습 흐름을 중심으로 훈련 플랫폼을 설계하였다. 사용자는 합성된 음성을 청취한 후 해당 상황이 보이스피싱에 해당하는지를 판단하며, 이후 제공되는 피드백을 통해 자신의 판단 결과를 확인한다. 이러한 구조는 고령자가 보이스피싱 상황에 대한 인식 능력을 점진적으로 향상시키는 데 목적을 둔다.

6.2 보이스피싱 시나리오 구성 원칙

훈련에 사용되는 시나리오는 실제 보이스피싱 범죄 사례를 기반으로 구성되었다. 기관 사칭, 가족 사칭, 금융 사기 등 대표적인 범죄 유형을 중심으로 시나리오를 분류하였으며, 실제 범죄에서 사용되는 발화 패턴을 반영하도록 설계하였다. 또한 시나리오는 향후 난이도 조절 및 확장이 가능하도록 구조화하여, 다양한 수준의 사용자가 단계적으로 훈련을 수행할 수 있도록 하였다.

6.3 사용자 인터랙션 설계

사용자는 시나리오 선택과 음성 생성 실행을 버튼 방식으로 수행하며, 필요한 경우 음성 파라미터를 슬라이더를 통해 조정할 수 있다. 이러한 설계는 자유 입력을 최소화하여 고령자의 조작 부담과 오류 발생 가능성을 줄이기 위한 것이다.

또한 음성 생성 과정에서는 현재 처리 단계와 상태를 직관적으로 인지할 수 있도록 시각적 피드백을 제공하여, 시스템 사용 과정에서의 혼란과 불안감을 최소화하였다.

Fig. 2에 제시된 음성 합성 설정 화면에서는 상단에서 보이스피싱 시나리오 유형과 문장을 선택하고 중단의 슬라이더를 통해서 안정성, 유사도 등의 합성 파라미터를 조절한 뒤 상단의 합성 버튼을 통해 오디오 플레이어로 결과를 확인할 수 있다.

[Fig. 2] Voice synthesis configuration screen

7. 시스템 구현 결과 및 동작 검증

본 장에서는 제안한 보이스피싱 예방 훈련 시스템이 설계 의도에 따라 정상적으로 동작하는지를 검증한다. 검증은 음성 데이터 수집, 모델 학습, 음성 합성, 훈련 콘텐츠 제공에 이르는 전 과정이 수행되는지와 소량 데이터 기반 음성합성 모델이 훈련 시나리오 전달에 필요한 최소한의 명료도를 확보하는지 확인하는 데 초점을 둔다. 모델 간 성능을 정량적으로 비교하는 대규모 벤치마크보다는 시스템 수준의 동작 가능성과 교육적 활용 가능성을 평가하는데 중점을 두었다.

7.1 검증 목적 및 검증 범위

본 검증의 목적은 제안한 시스템이 고령자 보이스피싱 예방 훈련 플랫폼으로서 기술적으로 구현 가능하며, 각 구성 요소가 상호 연동되어 정상적으로 동작하는지를 확인하는 것이다. 이에 따라 본 연구에서는 정량적 성능 비교 실험보다는 기능 단위의 동작 여부 검증을 중심으로 평가를 수행하였다.

검증 범위는 음성 데이터 수집 기능, 서버 기반 모델 학습 기능, 학습된 모델을 활용한 음성 합성 기능, 그리고 훈련 콘텐츠 제공 기능을 포함한다.

7.2 검증 시나리오 및 절차

시스템 검증은 실제 사용자 사용 흐름을 기반으로 한 시나리오를 통해 수행되었다. 사용자가 모바일 애플리케이션을 통해 음성을 녹음하면 해당 데이터는 서버로 전송되어 저장 및 전처리 과정을 거친다. 이후 사용자는 수집된 음성 데이터를 기반으로 음성합성 모델 학습을 요청하며, 서버에서는 해당 요청을 처리하여 모델 학습을 수행한다.

모델 학습이 완료된 이후, 사용자는 생성된 모델을 선택하여 보이스피싱 시나리오 텍스트에 대한 음성 합성을 수행하며, 생성된 음성은 훈련 콘텐츠로 제공된다. 본 연구에서는 이와 같은 음성 데이터 수집–학습–합성–콘텐츠 제공의 전 과정이 오류 없이 수행되는지를 단계별로 확인하였다.

7.3 기능별 동작 검증 결과

모델 학습 과정에서는 화자 1명당 평균 약 10~15분의 학습 시간이 소요되었으며, 이는 제안한 시스템이 소량의 음성 데이터 환경에서도 과도한 학습 비용 없이 모델 생성을 수행할 수 있음을 확인하는 데 활용되었다.

모델 학습 완료 후 생성된 음성합성 모델을 활용하여 보이스피싱 시나리오 텍스트에 대한 음성 합성을 수행한 결과, 모든 테스트 화자에 대해 음성 생성 기능이 정상적으로 동작함을 확인하였다. 본 검증에서는 음성 생성의 정량적 속도 측정보다는, 합성 음성이 정상적으로 생성되고 재생되는지 여부를 중심으로 확인하였다.

7.4 정량적 성능 평가 및 사용자 만족도 분석

본 연구는 제안한 음성합성 기반 훈련 시스템의 실효성을 객관적으로 검증하기 위하여 생성 음성의 명료도 평가와 실제 사용자 대상 만족도 조사를 수행하였다. 생성 음성의 명료도 평가는 단어 오류율(Word Error Rate, WER)을 기준으로 수행하였다. 본 연구의 목적은 방송 수준의 음질 확보가 아니라 보이스피싱 훈련 시나리오 전달의 정확성과 이해 가능성 확보에 있으므로 음성 인식 기반 지표를 활용하여 텍스트 전달 정확도를 분석하였다. Whisper Large-v3 모델을 이용하여 생성 음성을 자동 음성 인식 시스템에 입력하고, 인식된 텍스트와 원본 시나리오를 비교해 WER을 계산하였다. WER은 다음과 같이 정의된다.

$$WER = \frac{S + D + I}{N}$$

여기서 S는 대체 오류 수, D는 삭제 오류 수, I는 삽입 오류 수, N은 기준 문장의 전체 단어 수를 의미한다. 총 50명의 화자 데이터를 대상으로 분석한 결과 평균 WER은 6.46%로 측정되었으며 최저 2.86%, 최고 9.86%의 범위를 보였다. 이는 화자당 약 5분 분량의 소량 데이터 학습만으로도 훈련 시나리오 전달에 필요한 음성 명료도를 확보하기에 충분한 수준으로 판단된다.

또한 시스템의 교육적 활용 가능성을 검증하기 위하여 전남 지역 4개 기관에서 총 104명의 고령자를 대상으로 실증 교육을 실시하였다. 교육 종료 후 5점 리커트 척도 기반 설문 조사를 실시하였으며 '음성합성 체험 만족도' 및 '교육 전반 만족도' 항목에서 85% 이상의 참여자가 4점 이상으로 응답하였다. 이는 참여자가 자신의 음성이 인공지능 기반 음성으로 재현되는 과정을 직접 경험함으로써 보이스피싱 위험성을 보다 현실적으로 인식하였음을 보여준다.

이러한 정량적 분석 결과를 종합할 때 본 연구에서 제안한 소량 데이터 기반 화자 적응 전략은 보이스피싱 예방 훈련 시스템에 적용 가능한 수준의 음성 전달 정확도와 교육적 수용성을 확보한 것으로 판단된다.

8. 결론 및 향후 연구

본 연구는 고령자를 대상으로 한 보이스피싱 예방 교육의 실효성을 향상시키기 위하여 소량 음성 데이터 기반의 개인화 음성합성 모델과 모바일 기반 훈련 플랫폼을 설계 및 구현하였다.

정량적 성능 검증을 위해 생성 음성에 대한 단어 오류율 분석을 수행한 결과 평균 6.46%로 측정되었으며 이는 훈련 시나리오 전달에 필요한 음성 명료도가 확보되었음을 시사한다. 또한 104명을 대상으로 한 실증 교육 만족도 조사에서도 높은 수용성이 확인되었다. 이러한 결과는 제안한 소량 데이터 기반 화자 적응 전략이 실제 보이스피싱 예방 훈련 환경에서 적용 가능함을 시사한다.

다만 본 연구는 장기적 학습 효과에 대한 통제 실험이나 무작위 대조군 기반 비교 분석까지는 수행하지 못하였다. 또한 다양한 음성합성 모델 간 정량적 비교나 주관적 음질 평가(MOS) 실험 역시 포함하지 않았다. 향후 연구에서는 모델 간 성능 비교 및 장기적 교육 효과 분석을 통해 시스템의 효과성을 보다 체계적으로 검증하고 보이스피싱 음성 탐지 모델이나 대화형 인공지능 시스템과의 연계를 통해 더욱 지능적인 예방 훈련 시스템으로 확장할 필요가 있다.

REFERENCES

[1] L.S.Wijesinghe, "Features and Evolution of Phishing, Vishing, and Smishing," Master's Thesis, Lund University, 2025.

[2] L.Song, X.Zhang, and X.Zhou, "The Silence of the Phishers: Early-stage Voice Phishing Detection," Computers & Security, in press, 2025.

[3] D.Pehlivanoglu, A.Shoenfelt, Z.Hakim, A.Heemskerk, J.Zhen, M.Mosqueda, R.C.Wilson, M.Huentelman, M.D.Grilli, G.Turner, R.N.Spreng, and N.C.Ebner, "Phishing vulnerability compounded by older age, apolipoprotein E e4 genotype, and lower cognition," PNAS Nexus, Vol. 3, No. 8, pp.(Article page296), 2024.

[4] B.Zhang, H.Cui, V.Nguyen, and M.Whitty, "Audio Deepfake Detection: What Has Been Achieved and What Is Next," Sensors, Vol.25, No.7, pp.(Article 1989), 2025.

[5] M.U.Tanveer, K.Munir, M.Amjad, A.U.Rehman, and A.Bermak, "Unmasking the Fake: Machine Learning Approach for Deepfake Voice Detection," IEEE Access, Vol.12, pp.197442-197453, 2024.

[6] X.Chen, M.Sacre, G.Lenzini, S.Greiff, V.Distler, and A.Sergeeva, "The Effects of Group Discussion and Role-playing Training on Self-efficacy, Support-seeking, and Reporting Phishing Emails: Evidence from a Mixed-design Experiment," in Proceedings of the 2024 ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2024), ACM, 2024, pp.(Article 829).

[7] A.Baillon, J.deBruin, A.Emirmahmutoglu, E.vandeVeer, and B.vanDijk, "Informing, simulating experience, or both: A field experiment on phishing risks," PLOS ONE, Vol.14, No.12, pp.(Article e0224216), 2019.

[8] K.E.Sabo, J.Black, and D.M.Sarno, "Developing IMPAWSTER: Improving Meaningful Phishing Awareness With Simulated Training and Email Roleplay," Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol.67, No.1, pp.1665-1670, 2023.

[9] C.Zhou, W.Zhan, T.Huang, H.Zhao, and J.Kaner, "An empirical study on the collaborative usability of age-appropriate smart home interface design," Frontiers in Psychology, Vol.14, pp.(Article 1097834), 2023.

[10] E.Amouzadeh, I.Dianat, J.Faradmal, and M.Babamiri, "Optimizing mobile app design for older adults: systematic review of age-friendly design," Aging Clinical and Experimental Research, Vol.37, pp.(Article 248), 2025.

[11] E.J.Koh and H.W.Jung, "User Experience Design and Improvement Strategies for Chronic Disease Management Digital Healthcare Applications for Older Adults," The Journal of Transdisciplinary Studies, Vol.8, No.3, pp.349-368, 2024.

[12] S.Choi, S.Han, D.Kim, and S.Ha, "Attentron: Few-Shot Text-to-Speech Utilizing Attention-Based Variable-Length Embedding," in Proceedings of Interspeech 2020, 2020, pp.2007-2011.

[13] M.J.I.Basher, M.Kowsher, M.S.Islam, R.N.Nandi, N.J.Prottasha, M.H.Menon, T.A.Muntasir, S.A.Chowdhury, F.Alam, N.Yousefi, and O.Garibay, "BnTTS: Few-Shot Speaker Adaptation in Low-Resource Setting," in Findings of the Association for Computational Linguistics: NAACL 2025, 2025, pp.4971-4983.

[14] X.Liu, X.Ma, W.Song, Y.Zhang, and Y.Zhang, "High fidelity zero shot speaker adaptation in text to speech synthesis with denoising diffusion GAN," Scientific Reports, Vol.15, pp.(Article 6269), 2025.

[15] R.Uematsu, C.S.Leow, N.Kitaoka, and H.Nishizaki, "Improving Automatic Speech Recognition Model for Super-Elderly Voice Using Speech Synthesis Model," in Proceedings of the 2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, 2025, pp.986-991.

[16] Y.S.Kim, S.M.Lee, M.K.Choi, S.M.Jung, J.E.Sung, and Y.Lee, "The effects of speakers' age on temporal features of speech among healthy young, middle-aged, and older adults," Phonetics and Speech Sciences, Vol.14, No.3,

pp.37-47, 2022.

[17] W.Ping, K.Peng, A.Gibiansky, S.O.Arik, A.Kannan, S.Narang, J.Raiman, and J.Miller, "Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning," in Proceedings of the International Conference on Learning Representations (ICLR), 2018.

[18] E.Casanova, J.Weber, C.Shulby, E.Gölge, M.Hartmann, and G.Gelly, "YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for Everyone," in Proceedings of the 39th International Conference on Machine Learning (ICML), PMLR, Vol.162, pp.2709-2720, 2022.

정 유 태(Yu-Tae Jung) [준회원]

■ 2026년 2월 : 국립목포대학교 융합소프트웨어학과(공학사)

〈관심분야〉
컴퓨터 비전, 이미지 품질 평가, 이미지 재구성, 딥러닝 기반 음성 및 영상 처리

조 광 문(Kwangmoon Cho) [종신회원]

■ 1995년 8월 : 고려대학교 전산과학과(이학박사)
■ 1995년 9월 ~ 2000년 2월 : 삼성전자 통신연구소 선임연구원
■ 2000년 3월 ~ 2005년 2월 : 백석대학교 정보통신학부 교수
■ 2005년 3월 ~ 현재 : 국립목포대학교 컴퓨터학부 교수

〈관심분야〉
사물인터넷, 통신 소프트웨어, 인공지능 교육, 웹 서비스