https://doi.org/10.7840/kics.2022.47.10.1677

딥보이스를 악용한 보이스 피싱 피해방지 서비스 개발

김 소 운◆, 이 성 택°

Development of Voice Phishing Damage Prevention Service Misusing Deep Voice

Sowoon Kim◆, Sungtaek Lee°

요 약

딥보이스(Deep Voice)는 인공지능(AI)을 이용한 음성합성 기술로 온라인을 통하여 누구나 쉽게 음성합성 및 조작에 접근할 수 있는 기회를 제공하고 있어, 단순히 음성을 합성한 변조된 콘텐츠의 생산 뿐만 아니라, 해당 콘텐츠를 악용하여 사이버범죄 등이 확산되는 결과를 불러일으키고 있다. 유사범죄의 대표적인 사례가 보이스피싱이며, 해외에서는 물론 국내에서도 딥보이스를 악용한 보이스피싱 범죄가 나타나고 있어 이와 같은 피해를 줄일 수 있는 서비스가 개발되어야 한다. 피해의 대상자는 유명인, 실제 공공기관, 금융회사 등과 같이 전문직에서 근무하는 특정인 뿐만 아니라, 가까운 지인 혹은 가족 등을 사칭하는 맞춤형 피싱도 등장하고 있다. 본 연구에서는 전화통신 과정에서 발생할 수 있는 딥보이스 기술을 악용한 보이스피싱 피해를 방지하기 위한 서비스(프로그램)을 제안하려 한다. 해당 서비스는 딥러닝, 음성합성, 화자인식, 화자식별, 화자검증 등의 과정을 통하여 실제 화자와 딥보이스를 구분하는 목적으로 개발하고자 하며, 이는 단순 보이스피싱뿐만 아니라 다양한 범죄들의 위험으로부터 벗어날 수 있는 서비스의 확장으로 연계될 것이라고 기대한다.

키워드 : 딥보이스, 음성합성, 화자인식, 화자식별, 화자검증
Key Words : Deep Voice, Voice Synthesis, Speaker Recognition, Speaker Identification, Speaker Verification

ABSTRACT

Deep Voice is a speech synthesis technology using AI that provides an opportunity for anyone to easily access speech synthesis and manipulation through online. This causes not only the production of altered content that synthesizes voice, but also the spread of cybercrime by misusing the content. A representative example of a similar crime is voice phishing, and there are appearing voice phishing crimes that misuse deep voice abroad as well as in Korea, so a service that can reduce damage should be developed. The victims are not only celebrities, public institutions, financial companies, etc., but also acquaintances or family members. In this study, we are going to propose a service (program) to prevent voice phishing damage that may occur in the course of telephony misusing deep voice technology. This service intend to develop this service for the purpose of distinguishing the actual speaker from the deep voice through the processes of deep learning, voice synthesis, speaker recognition, speaker identification, and speaker verification. It is expected that it will be linked to the expansion of services that can avoid the risk of various crimes as well as simple voice phishing.

◆ First Author : Yong In University Department of Computer Science, sounso0805@naver.com, 학생회원
° Corresponding Author : Yong In University Department of AI, totona22@yiu.ac.kr, 정회원
논문번호 : 202208-182-0-SE, Received August 15, 2022; Revised August 25, 2022; Accepted August 25, 2022

Ⅰ. 서 론

인공지능이 발달하면서, 전 세계적으로 정치인, 연예인처럼 공인들의 동영상 합성에 사용되는 인공지능 기술기반의 딥페이크(Deepfakes)는 사회에 점점 많은 문제들을 불러일으키고 있다.[1] 이러한 딥페이크 기술과 함께 딥보이스 기술의 악용 사례들도 등장하고 있다. 딥보이스 기술이란 인공지능을 이용한 음성합성 기술로, 단순히 음성을 위조한 콘텐츠 뿐만 아니라, 여러 유형의 영상물이 제작되는 콘텐츠들도 다양하게 나타나고 있다. 인터넷 상에서 '딥보이스 목소리, 딥보이스 만들기, 인공지능 음성 만들기' 등을 검색하면 쉽게 음성 조작이 가능한 다양한 솔루션 들을 접할 수가 있다. 인공지능 기술의 발전으로 인해서 누구나 손쉽게 음성 조작이 가능해진 만큼 앞으로 다양한 유형의 범죄가 증가할 가능성도 높아질 것으로 예측할 수 있을 것이다.

보이스피싱(Voice Phishing) 사기 피해는 최근 10여년간 빠르게 증가하면서 소비자들의 금융자산의 안전을 위협하고 있다. 보이스피싱은 피싱(phishing)에서 파생된 용어이며, 피싱사기란 '전화 등을 통해 개인정보를 낚아올린다'는 의미에 개인정보(Private Data)와 낚시(Fishing)을 합성한 신조어이고, 보이스피싱은 피싱이라는 단어 앞에 음성(voice)의 의미가 더해진 것이다. 즉 보이스피싱은 전자통신수단 중 특히 전화를 통해 불법적으로 개인정보를 이용해 금전적 이득을 꾀하는 범죄를 의미한다(금융감독원, 2020).[2]

최근 인공지능 기반의 음성합성 및 음성조작 기술 등이 발전함에 따라 보다 지능화되는 보이스 피싱 범죄로 인하여 금융사기 피해가 지속적으로 증대하고 있다. 초기의 보이스 피싱은 현재와 다르게 발음이 어눌하고 음질도 좋지 않았지만, 과거와 달리 최근의 보이스 피싱은 세련된 말투와 전문적인 용어를 구사하여 사람들을 더욱 혼란스럽게 하고 있다. 뿐만 아니라, 과거에는 판단력이 부족한 노년층을 대상으로 하는 보이스 피싱 피해가 많았으나, 최근에는 세련된 말투와 전문직에서 사용하는 전문용어를 사용하며 젊은 세대(20~30대)들도 많은 피해를 당하고 있는 것으로 나타났다.[3]

1.1 딥페이크 사례

인공지능 기술로 얼굴과 음성을 변조하는 딥페이크 기술을 악용한 피싱 범죄에 대한 우려와 경고의 목소리가 높아지고 있다. 인물의 이미지를 업로드하면 자동으로 여성의 신체에 그 인물의 얼굴을 합성해주는 프로그램이다. 네덜란드 AI 연구소 센서티(Sensity)(舊 딥트레이스(Deeptrace))의 보고서에 따르면, 2020년 7월까지 텔레그램에서 전세계 10만 4582명의 여성이 '딥페이크봇'에 걸려 자신의 얼굴이 알 수 없는 사람의 신체와 합성되는 피해를 당한 것으로 보고하고 있다. 이 중 약 70%는 유명인이 아닌 일반인들의 개인 SNS 등에서 합법적인 동의없이 수집한 이미지와 영상인 것으로 조사되었다.

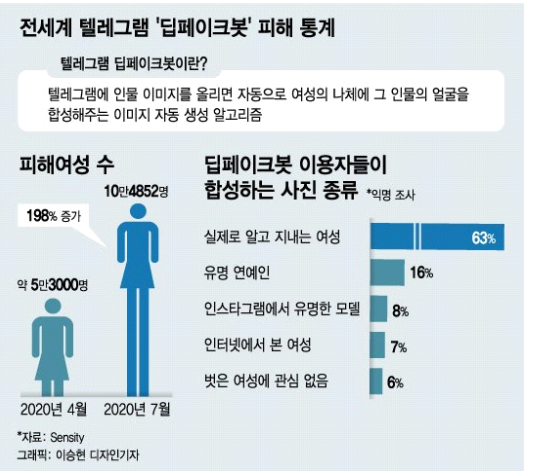

그림 1. 전세계 텔레그램 '딥페이크봇' 피해 통계 자료
Fig. 1. Global Telegram 'DeepFakebot' Damage Statistics

1.2 딥보이스 사례

포브스에 의해 밝혀진 법원 문서들에 따르면, 은행 임원을 사칭하여 딥보이스 기술을 악용해 보이스피싱 전화를 걸어 자신이 곧 회사를 인수할 것이며 그 과정에 돈이 필요하다며 돈을 요구한 내용을 확인할 수 있다. 임원과 친분이 있었던 은행장은 목소리가 너무 동일하였기에 임원의 요구에 응하게 되었는데, 이 사건으로 돈은 범죄자들의 손에 들어가게 되었고, 이 돈은 미국 소재 은행까지 연루되어 미국의 수사관과 두바이 수사관의 공조까지 이어지는 번거로운 일이 발생했다. 앞서 언급한 것처럼 해외에서는 딥보이스에 대한 사례가 등장하고 있으며, 국내에서도 딥보이스를 악용한 보이스피싱이 등장하여 향후 유사한 사례가 증대될 것으로 예상할 수 있다. 국회입법조사처에서 발간한 <주요국의 피싱(Phishing) 사기 입법 · 정책 동향과 시사점> 보고서[18]에 따르면, 범죄자들이 유명인, 실제 공공기관, 금융회사에 근무하는 특정인들을 대상으로 하는 피해부터 가까운 지인 혹은 가족 등을 사칭하는 맞춤형 피싱도 등장할 수 있다는 우려의 목

소리가 나오고 있다.

1.3 딥페이크와 딥보이스 피해유형

[표 1]에서 보는 것과 같이 딥보이스와 딥페이크를 악용한 사례들은 다양한 유형을 통해서 나타나고 있다. 주로 영상 또는 이미지와 관련된 악용사례들이 상대적으로 더 많이 발생하지만, 딥보이스를 악용한 사례도 나타나고 있음을 확인할 수 있다. [표 2]에서는 딥페이크 성적 허위영상 처리 현황에 대한 자료로 심의(차단) 되거나 자율규제(삭제)한 딥페이크 성적 허위영상 수가 크게 늘어났다는 사실을 확인할 수 있다.

현재 다양한 자료들을 기반으로 조사한 결과, 딥페이크를 악용한 범죄 및 피해사실들은 다양하게 나타나고 있지만, 딥보이스를 악용한 인한 피해사례들은 상대적으로 많이 확인되고 있지 않은 상황이다. 하지만, 인공지능 기술의 지속적인 발전으로 인하여 딥보이스를 악용한 피해사례들도 증가할 것으로 예측 가능하며, 특히 딥보이스를 악용한 보이스피싱 사례들이 많이 등장할 것으로 예상되어 본 연구에서는 딥보이스를 악용한 보이스피싱 피해를 예방하고자 음성합성, 화자인식, 화자식별, 화자검증 등의 기술을 기반으로 피해예방 서비스를 제안하고자 한다.

표 1. 딥보이스와 딥페이크 피해
Table 1. Deep Voice and Deep Fake Damage

Technology	Damage Pattern	Type
Deep Voice	Voice Phishing Damage	Voice
Video Clip	Video Creation and Distribution	Video
DeepFake Bot	Face Synthesis Damage	Image
Impersonation of Another Person	Fake News Generation	Text/ Video
Defamation of Character	Use a Celebrity or Other Person's Face to Blame Others	Image

표 2. 딥페이크 성적 허위영상 처리 현황[17]
Table 2. Deep Fake Sexual False Image Processing Status

Sortation	Deliberation (blocking)	Self-regulation	Sum
2020. 6~7	473	75	548
2021. 1~9	537	871	1,408

Ⅱ. 보이스피싱 피해사례

2.1 보이스피싱 피해

보이스피싱 피해 사례 및 범죄 조직의 분석 결과에 따르면 보이스피싱 범죄는 인터넷 전화, 국제전화, 선불 휴대폰, 대포 통장 등과 같이 정보통신기술과 금융시스템을 활용한 국제적이자 국가간의 범죄라고 할 수 있다. 보이스피싱 범죄단은 구체적으로 역할이 분화되어 있음과 동시에 조직적으로 범행을 하고 있고, 사칭하는 기관들이 다양하기 때문에 피해자들이 계속해서 피해를 받고 있는 것으로 드러났다. 또한 특정 지역, 직업, 연령, 성별에 관계없이 무차별적으로 피해가 발생하는 것으로 분석되었다. 특히 피해자들이 주로 노인층일 것이라는 일반적인 가정들과 다르게 평균 연령은 50세인 것으로 나타났다.[14]

2.2 보이스 피싱의 진화(딥보이스 악용)

보이스피싱은 전기통신을 범행도구로 이용하는 범죄에서 과학기술의 발달로 다양한 기술을 범행도구로 사용하며 진화를 거듭해왔다. 최근 인공지능 기술을 접목한 딥페이크(Deepfake)와 딥보이스(Deepvoice) 등 새로운 기술로 지속적인 진화를 거듭하며 피해자의 희망과 공포를 이용하여 피해자들의 돈을 갈취하고 있다. 그 결과 2006년 최초로 발생한 이후 꾸준히 증가하여 지난 2020년 범죄 피해금액은 약 7,000억원으로 확인되었다. 지금까지 정부의 다양한 노력에도 불구하고 보이스피싱의 피해는 지금도 발생하고 있다.[15]

최근 디지털 기술의 발달은 보이스 피싱 범죄의 고도화에도 영향을 미치고 있다. 인공지능 기술을 활용하여 딥페이크 형태의 음성변조나, SIM 박스를 악용한 발신번호 변경 등을 악용한다. 또한 피해자의 기기에 악성 코드를 심어 전화를 탈취하는 스미싱도 성행하고 있다. 이에 따라 음성인식, 텍스트 분석, 발신번호 변작 차단, SIM 박스 패킷 분석 및 전파 탐지, 악성앱 탐지 등과 같은 기술적 대응책들이 연구되고 있다. 현재보다 더 많은 연구개발과 피해방지 서비스개발이 이루어져야 하겠지만 대응제도는 개선이 미흡한 상황이다. 특히 형사사법공조체제는 제대로 기능하지 않는 상황이기 때문에 실질적인 수사가 어렵고, 보이스피싱 범죄에 대한 관리 및 처벌과 제재가 약하면서 범사회적 차원의 협력과 피해자 구제를 위한 사회적 완충 조치들도 부족한 상황이다.[16]

Ⅲ. 딥보이스 관련기술

3.1 음성합성 기술

음성이란 인간이 지니고 있는 기본적인 능력들 중에서 가장 중요한 요소 중 하나이다. 우리가 속박감을 거의 느끼지 않으면서 자유롭게 구사할 수 있는 자연스러움과 동시에 효과적인 정보교류의 수단이라고 할 수 있다. 더 나아가 음성에 의해 표현되는 말은 인간과 인간과의 사이에서의 의사소통 수단 뿐만 아니라 논리적으로 사물을 생각하는 상황에서도 중요한 역할을 한다.[4]

음성합성 시스템은 텍스트 데이터를 입력하면 그에 대응하는 음성을 출력해주는 시스템이다. 주어진 음성을 대응하는 텍스트로 전사해주는 음성인식과 대비되는 시스템으로, 최근 인공지능 비서, 오디오북, 인공지능 스피커 등에 대한 관심 증가와 함께 주목을 받고 있다.[5]

음성합성 기술의 기본 원리는 문자를 음성으로 변환하는 기술인 문자 음성 자동변환(TTS, Text to Speech)기술이다. 역사적으로 다른 음성관련 기술보다 가장 선행적으로 연구된 기술이다. 초기 음성합성에 관한 연구는 인간의 발성기관에 대해 모사였으나, 근래에는 문자 음성 자동변환(TTS) 시스템으로 구현되면서 원하는 정보를 음성으로 전달할 수 있어 다양한 분야에 이용되고 있다. 음성합성의 방식이 과거의 제한적인 단어와 문장으로 구성된 편집 합성 방식을 뛰어넘어 무제한 음성합성 방식인 규칙 합성 방식(임의의 단어와 문장을 표현 가능한 입력 기호열로 하며, 이에 대응하는 음성을 출력하는 음성합성 방식)을 도입하고 있고, 연결 합성 방식(자주 사용되는 문장을 녹음하여 음소 단위로 나눔 뒤에 필요한 문장으로 조합하는 음성합성 방식)이 개선되고 있다.[6]

최근 음성합성 기술은 딥러닝 학습과 연계되어 사람의 목소리를 학습하면서 문맥에 따라서 강세, 높낮이, 발음하는 방법 등을 습득하는 단계로 발전하였다.

3.2 딥러닝 기술

딥러닝이란 사람의 신경세포를 모사하여 기계가 학습하도록 하는 인공신경망(Artificial Neural Network) 기반의 기계 학습법이다. 최근 이미지 인식, 자연어 처리, 음성 인식 등과 같이 다양한 분야의 발전에 기여하고 있으며[7], 음성합성 기술이 기존에 존재했던 기술임에도 불구하고 최근 과거와 다르게 더욱 자연스러운 기계음을 낼 수 있게 된 것은 딥러닝 기술의 발전에 따른 학습역량의 향상에 기인한 것이라 할 수 있다. 이와 같은 딥러닝 기술은 대량의 데이터를 학습할 수 있는 하드웨어 기술의 발전과 더불어 학습할 수 있는 대량의 목소리를 기반으로 문맥에 따라 강세, 높낮이, 발음 등 다양하게 발음하는 방법을 학습하게 된다.

딥러닝 알고리즘 중에서 순환신경망(RNN, Recurrent Neural Network) 구조는 음성/오디오 분석에 사용되는 기본적인 알고리즘으로 영상처리와 다르게 원천 데이터의 형식이 일반적으로 1차원 데이터라는 특성과 시계열적이라는 특성을 적용하기에 적합하다고 할 수 있다. 또한 LSTM (Long Short‐Term Memory)은 기울기 소실문제(gradient vanishing)을 방지하기 위해서 제안된 알고리즘으로써, 1997년 제안된 구조로 현재는 기본 구조 외에도 다양하게 변형된 LSTM 연구가 많이 진행되고 있으며, RNN과 다른 점은 은닉층(hidden layer) 내에 메모리 기능을 넣음과 동시에 메모리를 조절(쓰기, 지우기, 출력하기)할 수 있는 훈련을 통하여 결과값을 도출하는 것이 기본적인 특성이다. LSTM은 기존의 RNN과 비교했을 때 장기종속성(Long-term Dependencies) 문제에 특화되어 있어, 학습하는데 많은 시간과 데이터가 필요하지만, 많은 양의 데이터가 확보된 상황에서는 보다 뛰어난 성능을 보이는 것으로 검증되었다.[8]

이와 같은 딥러닝 알고리즘들을 활용하여 다량의 음성데이터 학습을 통한 음성합성기술의 발전은 화자인식, 화자식별, 화자검증 과정의 정확성을 개선해야만 가능해 질 수 있다.

3.3 화자 인식 영역

화자인식은 음성인식과 가장 밀접한 관계가 있다. 말하는 사람, 즉 화자의 음성파에 포함되어 있는 개인적인 정보를 이용하여 누구의 음성인지를 자동으로 판정하는 기술을 의미한다. 화자인식은 화자독립 음성인식기술의 원리와 매우 유사하다. 넓은 의미에서의 화자인식은 사람이 음성을 청취하고, 그 스펙트럼을 이용하여 화자를 식별하는 원리를 연구하는 분야를 포함한다.

화자인식은 발성방법에 따라서 문맥종속형과 문맥독립형으로 나눌 수 있다. 문맥종속형은 미리 정해진 문장을 발성하게 하는 반면에 문맥독립형은 특별히 지정하지 않고 발성하도록 하는 것이다. 일반적으로 음성은 음향학적, 음성학적으로는 변화가 많다. 때문에, 문맥독립형은 문맥종속형에 비해서 더욱 더 많은 훈련 데이터를 필요로 한다. 문맥종속형은 일부 시스템에서 비음과 같이 특별한 음소를 이용하는 경우도 있다. 하지만 대부분 단어(이름, 핵심어, ID 등) 혹은

임의로 선정한 문장을 사용한다. 문장을 사용하는 경우에는, 독립단어를 이용하는 경우에 비해서, 문장 가운데의 단어발성 혹은 문장발성이 화자에 따라 달라지는 예시가 많기 때문에 인식률 향상이 나타날 수 있다.

화자인식에서 화자가 협조적인가 아닌지에 따라서 인식률이 크게 달라진다는 점이 어려운 부분이다. 문맥종속 시스템 및 문맥독립 시스템에서 가장 큰 단점은 등록된 화자의 음성을 통해 발성된 핵심어 또는 지정한 문장을 녹음기를 이용하여 녹음한 후, 마이크로 입력할 경우에 등록된 화자로 승인된다는 점이다. 이 문제를 해결하기 위해서 핵심어들로 구성된 소규모의 단어세트를 이용하여 이들을 랜덤하게 제시하여 발성하게 하는 방법 등이 활용되고 있다.

화자인식 시스템은 응용방법에 따라 화자식별과 화자검증 시스템으로 나뉜다.[9]

3.3.1 화자식별

화자식별은 고립단어 인식과정과 유사하게 등록된 표준패턴 중 어떤 화자의 패턴과 입력 음성의 패턴 유사도가 가장 높은지를 비교하여 화자를 결정하는 것을 말한다. 범죄수사의 경우, 현장에서 녹음된 범죄자의 음성과 혐의자의 음성을 비교하여 범죄여부를 판단한다. 기업에서는 직원들의 음성을 녹음하여 등록함으로써 출퇴근 관리에 활용한다. 이와 같은 과정에서 화자식별은 화자검증과 화자식별과정이 결합되어 이루어진다.[9] 즉, 시스템에 음성을 입력한 뒤에 등록된 화자 모델을 검색하여 음성과 가장 일치하는 화자를 찾아주는 기술이다. 화자식별 기술의 가장 큰 단점은 시스템에 등록되지 않은 화자일 경우에, 음성을 구분하지 못하고 등록된 음성들 중 가장 유사한 화자로 인식한다는 것이다.

I-vector는 현재 화자인식 분야와 언어 인식 분야에서 가장 널리 사용되는 기법 중 하나이다. 음성이 가지고 있는 다양한 변이성을 낮은 차원의 고정된 크기의 벡터로 표현할 수 있다는 장점이 있다.[10] I-vector는 음성이 내포하는 화자와 채널 종속 공간을 개별적으로 처리하는 방식의 JFA[1](모델 GMM[2])의 스피커 및 세션 가변성 모델[11])의 연장이며, 화자와 채널 변동성 모두 모델링한 전체변이성 공간(total variability space)의 특징 벡터로 표현될 수 있다. 화자인식 시스템의 특징추출 방법인 I-vector 기법과 화자의 특성을 이와 결합시켜서 심층신경망(DNN, Deep Neural Network) 기반의 화자 식별 문제에 적용한 접근 방식을 제안하고 문성환(2018)의 연구에서 실험한 결과 [표 3]에서 알 수 있듯이 화자 특성을 결합한 모델이 baseline의 성능보다 뛰어넘는 사실을 확인할 수 있다. 그 중에서도 화자의 나이(age)를 결합시킨 결과가 성능에 큰 영향을 미치는 사실을 확인할 수 있다. 대체적으로 세분화 되어있으며, 화자마다 균등하게 분포되어있는 특성일수록 높은 성능을 보이는 것을 확인할 수 있다. 이러한 결과에 따라서 화자 특성 4가지 모두를 결합시켜서 더욱 구체적인 특성을 표현하는 모델의 성능이 가장 뛰어나는 것을 확인할 수 있다.

표 3. 화자식별 결과 정확도[12]
Table 3. Speaker Identification Results Accuracy

Model [Dimension]	Accuracy [Correct / Total]
I-vector(baseline) [200]	90.67% [408/450]
I-vector + gender [202]	92.00% [414/450]
I-vector + region [208]	93.11% [419/450]
I-vector + age [206]	94.00% [423/450]
I-vector + race [203]	91.33% [411/450]
I-vector + gender + region + age + race [219]	97.78% [440/450]

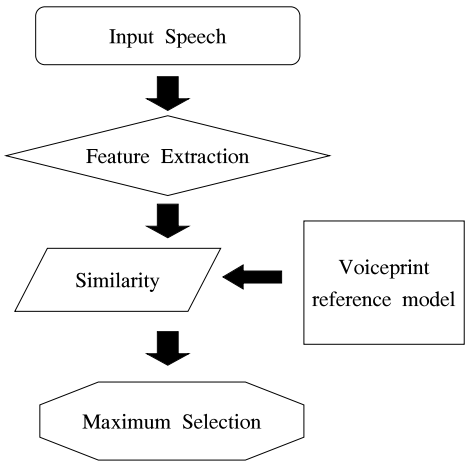

그림 2. 화자 식별 방식
Fig 2. Speaker identification method

1) JFA : Joint Factor Analysis
2) GMM : Gaussian 분포가 여러개로 혼합되어 있는 clustering 알고리즘

TIMIT[3] 데이터(딥러닝 초창기 평가를 위한 일반적인 데이터로, 미국의 8가지 방언을 사용하는 630명의 화자로 구성되어 있으며, 한 사람이 읽은 10개의 문장으로 이루어져 있다.)에서 실험한 결과, I-vector와 화자의 특성을 결합시킨 모델이 기존 결과의 성능을 넘어섰고 모든 화자의 특성을 결합한 모델의 결과가 가장 좋은 성능을 보인다는 사실을 확인할 수 있었다. 사전 화자의 특성 정보가 확보되는 상황에서는 발화된 화자의 음성과 특성을 결합하여 활용함으로써 고성능의 화자인식을 가능하게 될 것이라는 연구 결과가 나왔다.[12]

[그림 2]는 화자 모델을 등록한 시스템의 화자 식별 과정을 나타낸 그림이다. 시스템에 음성 파일을 저장한 후, 특징을 추출한다. 등록된 화자 모델을 검색하여 화자 참조 모델(Voiceprint reference model)과 유사도가 높은 화자를 찾는다.

3.3.2 화자검증

화자검증 시스템은 저장된 화자의 음성과 입력된 음성 사이의 유사도를 구하여, 저장되지 않은 목소리의 화자를 의미하는 '사칭자 모델'과 유사도 간 비율을 측정한다. 사칭자 대비 신뢰할 수 있는 비율에 해당하는 기준값에 따라서 화자 일치 여부를 검증하는 것이다.[19]

검증을 요구하는 화자의 발성과 그 화자의 등록된 기준 패턴을 비교하여 미리 정해놓은 임계값(발생 확률값)을 넘어서면 승인결과를 출력하고, 그렇지 않으면 거절결과를 출력한다. 음성을 여러 서비스에 응용할 수 있다.[9]

이를 통해서 시스템은 등록되지 않은 목소리를 구별하고, 딥페이크 탐지기술의 경우에는 avi 또는 mp4 형식의 딥페이크 활용이 의심되는 동영상에서 개별 프레임으로 잘라내어 분석하고자 하는 프레임을 이미지로 변환 후에 탐지한다.

문장 독립 화자 검증 분야에서 가장 널리 이용되고 있는 심층 화자 임베딩 방식을 발전시키기 위해서 그룹기반 화자 임베딩 방식이 정영문(2021)에 의해 연구되었다. 기존에 있는 심층 화자 임베딩에 화자의 그룹에 관한 정보를 담은 그룹 임베딩 벡터를 도입함으로써, 화자 임베딩이 나타낼 수 있는 전체 화자의 검색 공간을 줄여, 기존의 심층 화자 임베딩 방식을 향상시키는 것이다. 정영문(2021)의 연구에서 Ablation study(특정 구성 요소를 제거하여 AI 시스템의 성능을 연구하여 시스템 전체에 대한 구성 요소의 기여도를 파악)를 통해서 제안한 방식의 효용성을 확인할 수 있었다. 또한, A-Softmax(A-Softmax 비용 함수는 화자 임베딩 벡터 사이의 각도에 마진을 적용함으로써, 화자 임베딩 벡터 사이의 각도를 확대시켜주는 역할을 함)를 기반으로 한 비용 함수로 기존에 제안된 방식들보다 더 높은 성능을 달성하였다.[13]

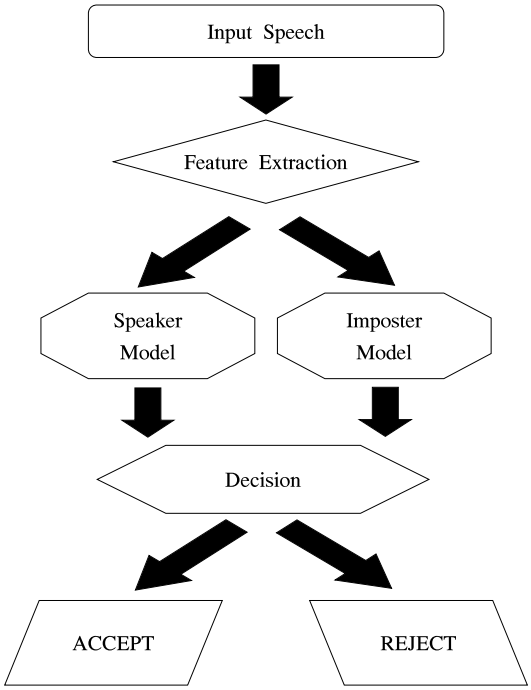

그림 3. 화자검증 방식
Fig. 3. Speaker Verification method

[그림 3]은 화자검증 방식을 나타내는 것으로, 음성파일을 저장 후에 특징을 추출하고, 화자모델(Speaker model)과 사칭자 모델(Imposter model)간의 유사도 간 비율을 측정하게 된다. 사칭자에 대비해 신뢰할 수 있는 비율에 해당하는 기준값에 따라서 화자일치 여부를 검증하게 된다.

Ⅳ. 서비스 개발

4.1 서비스 제안

본 연구의 목적은 딥보이스를 악용한 보이스피싱을 예방하기 위함이다. 딥보이스를 악용한 보이스피싱 전화를 받았을 경우에, 10단어 이상의 음성을 듣고 사람의 목소리인지, 딥보이스를 사용한 음성인지를 판단해주는 프로그램을 제안하고자 한다.[12] 앞에서 언급한 화자식별 기술을 적용하기 위하여 사전에 프로그램에

3) TIMIT : TIMIT는 다양한 성별과 방언을 가진 미국 영어 사용자의 음성 및 어휘 전사 연설 모음입니다.

본인의 음성을 미리 저장하여, 해당 프로그램에서 사용자와 발신자의 구별을 위한 과정의 비중을 줄여, 프로그램의 성능을 향상시키고자 한다.

프로그램 안에서 적용된 기술인 일반 음성과 다르게, 주파수 및 음역대에서 미세한 차이를 보이는 딥보이스 음성을 10단어 이상 듣게 될 시에 탐색하게 되는 기능을 탑재한다.

4.1.1 서비스(프로그램) 작동 방식

[그림 4]는 본 연구에서 제안하는 서비스(프로그램) 작동 방식을 그림으로 나타낸 것이다. 통화가 시작되면 프로그램이 자동으로 실행된다. 10단어 이상의 음성을 들은 후, 발신자가 딥보이스라고 판단되면 경고알람이 작동된다. 사전에 전화를 자동으로 종료되게 하거나 수동으로 종료할 수 있도록 설정이 가능하게 한다.

딥보이스를 악용한 보이스피싱 전화라면 전화를 종료한 후에 신고하기 기능이 뜨게 되고, 바로 신고가 가능할 수 있도록 연계한다.

만약, 딥보이스를 악용한 보이스 피싱이 아닐 경우에는 프로그램을 수동으로 종료한 후 전화는 계속 유지할 수 있도록 한다. 더 나아가 기본적인 표준어 뿐만 아니라 각 지역에서 사용하는 방언, 신조어, 합성어 등과 같이 현존하고 있는 많은 언어들의 학습을 통해 서비스로부터 소외될 수 있는 지역 및 세대를 최소화하는 서비스로 개발되어야 할 것이다.

4.1.2 서비스(프로그램)의 사용의 다양성

본 연구에서 제안하는 프로그램은 보이스피싱 피해방지에서 끝나는 것이 아니라, 음성 파일들을 악용한 다양한 문제들을 방지하기 위해 딥보이스가 사용된 음성이라고 의심되는 음성 파일을 분별해주는 기능도 제안하고자 한다. 딥보이스가 사용된 음성 파일일 경우에는 사용자에게 경고메세지 및 신고관련 기능을 제공하며, 해당 음성 파일의 경로 등과 같은 세부 정보들을 기반으로 악용의도를 차단하게 한다.

본 서비스의 고도화를 위하여, 개인정보 활용 및 이용동의자들의 통화내역을 녹음하여, DB에 지속적인 딥보이스 악용데이터의 수집을 지속하고, 해당 데이터를 기반으로 지속적인 학습을 실시하고자 한다. 지속적인 학습의 결과는 서비스 내에서 딥보이스를 사용한 음성 판별의 정확성을 높일 수 있을 것이라고 기대한다. 뿐만 아니라, 딥보이스가 악용된 사례 데이터를 수집하여, 해당 사례의 딥보이스 악용 진위여부를 판단할 수 있도록 전문가시스템을 활용하여, 딥보이스를 악용한 보이스피싱 예방의 성과를 높이고자 한다.

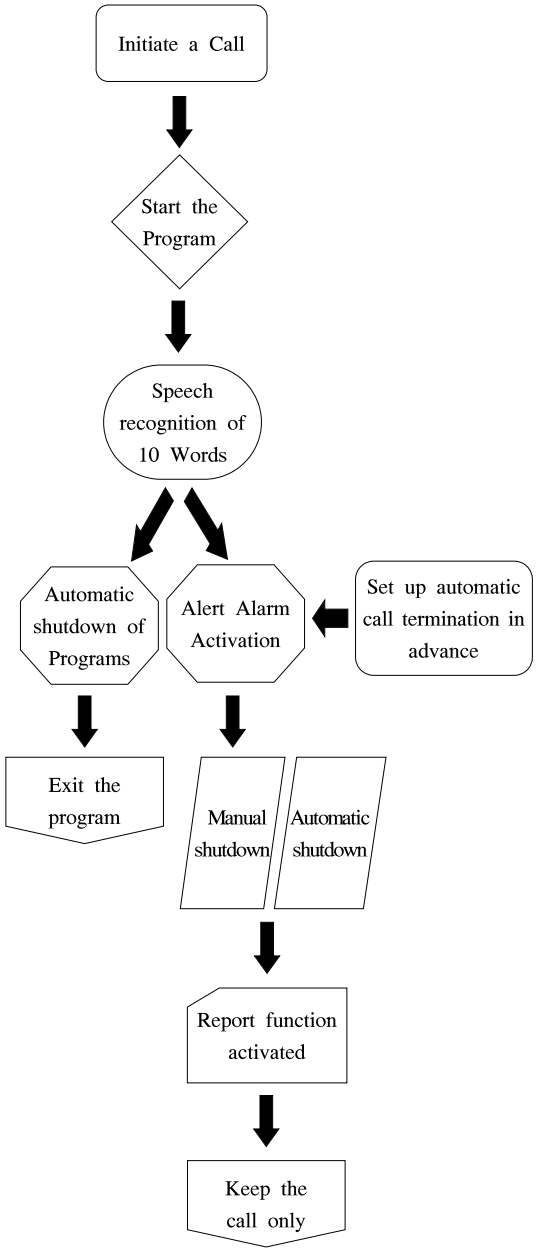

그림 4. 프로그램 동작 절차
Fig. 4. Program operating procedures

4.2 기대효과

본 서비스는 개인의 사용과 공공기관의 사용으로 크게 구분할 수 있다. 첫째, 개인이 사용할 때에는 개인이 사용하는 휴대폰에 설치하여 사용할 수 있다. 상황 판단을 스스로 하기 힘든 유년층과 노년층에게는

큰 도움이 될 것으로 기대한다. 딥보이스를 악용한 보이스피싱 전화를 받았을 때, 단번에 눈치를 채는 사람도 있을 것이다. 하지만 그 상황을 마주하게 된다면 혼란스러운 감정으로 제대로 대처하기 어려운 사람도 있을 것이다. 그와 같은 상황에서 해당 서비스의 도움을 받는다면 딥보이스를 악용한 보이스피싱으로부터 보다 안전해질 수 있을 것이다. 이처럼 일상생활 속에서뿐 아니라, 각종 공공기관에서 활용이 가능하다. 소방서, 경찰서와 같이 전화 한 통 한 통이 중요한 기관에 딥보이스를 악용한 보이스피싱으로 인한 피해가 발생하지 않도록 본 서비스의 도입은 꼭 필요한 것이라 할 수 있을 것이다.

국내에서 딥페이크 기술을 악용한 영상물들이 여러 유형으로 제작되어 범죄에 쓰이고 있고, 점차 피싱에 사용될 가능성이 높다는 지적이 나오고 있다. 본 연구에서 제안하는 서비스(프로그램)는 피싱의 직접적인 방지 뿐만 아니라, 다양한 범죄들의 위험으로부터 예방할 수 있는 효과를 제공할 것이라 기대한다. 이는 각종 사이버범죄로부터 조금 더 안전한 공동체 사회를 만들어 갈 수 있을 것이라고 기대한다.

V. 결 론

본 연구에서는 인공지능을 이용한 음성합성 기술인 딥보이스 기술을 악용한 보이스피싱을 예방하기 위한 서비스를 제안한다.

앞서 보이스피싱으로 인한 문제들에 대해서 살펴보았다. 보이스피싱에 대한 피해는 오랜 시간동안 끊임없이 일어나고 있다. 다양한 수법을 사용하여 다양한 연령층들을 혼란스럽게 하며, 이는 단순한 보이스피싱임에도 불구하고 예방하기에 어려움이 있을 수밖에 없다. 더 나아가, 인공지능 기술이 발전하면서 이를 악용한 범죄들을 살펴보았다. 해외에서는 지인을 사칭하여 딥보이스를 악용한 보이스 피싱 전화를 받아서 큰 피해를 본 사례도 볼 수 있었다.

이와 같은 피해를 줄이기 위해서 화자인식, 화자식별, 화자검증 등의 기술을 활용한 서비스를 제안한다. 통화연결과 동시에 서비스(프로그램)가 실행되고, 발신자가 딥보이스를 악용한 보이스피싱인지 아닌지 판단할 수 있는 서비스라고 할 수 있다. 딥보이스를 악용한 발신자라고 판단되면, 신고를 할 수 있는 서비스를 제공하며, 보다 정확한 구별이 가능하도록 서비스를 고도화하기 위하여, 사전 동의자들의 통화내역들과 사례들을 수집하여, 지속적인 학습을 한다면 피해 예방을 더욱 강력하게 할 수 있을 것이라 기대한다.

본 연구에서 제안하는 서비스(프로그램)는 피해 사건에 연루된 사람들의 시간과 수고로움을 덜 수 있을 것이며, 보이스피싱을 비롯해 각종 범죄를 예방하고 사람들의 자산을 안전하게 지키는 것에 있어서 분명하게 도움이 될 것이라고 기대한다. 개인 뿐만 아니라, 기관 및 기업에서도 본 서비스를 역으로 도입한다면, 소비자들이 불편해하던 홍보성 마케팅 통화 등에 민감한 소비자들로부터 기피대상으로부터 벗어나 건전한 고객관리가 가능할 것으로 기대한다.

현재 I-vector 기법과 화자의 특성을 결합한 모델의 결과가 가장 좋은 성능을 보인다는 화자식별에 대한 연구 결과와, A-Softmax를 기반으로 한 비용 함수로 기존에 제안된 방식들보다 더 높은 성능을 보인 화자검증에 대한 기술들보다 높은 정확도를 위해서는 많은 연구가 필요할 것이라고 생각한다.

본 연구에서 다루지 못한 개인정보 문제와 서비스를 사업화하기 위한 비즈니스 모델과 사업성 연구가 추가적으로 필요할 것으로 판단된다.

References

[1] J. Kim, et al., "Investigate the latest technology trends in data-driven deepfake detection techniques," *J. Inf. Protection (KIISC)*, vol. 30, no. 5, pp. 79-92, Oct. 2010.

[2] M. Kim and E. Kim, "Analysis of voice phishing damage experiences and influencing factors," *Korea Consumer Agency*, vol. 52, no. 1, pp. 52-71, Dec. 2021. (http://doi.org/10.15723/jcps.52.1.202104.53)

[3] B.-J. Lee, et al., "Features identification of 'Recently voice fishing' by applying ICT technologies," *2016 Korea Commun. Assoc. Winter Conf. Papers*, pp. 360-361, Gangwon-do, Korea, Jan. 2016.

[4] I. Yoo, "Overview and activating plan for voice recognition business in Korea," The Graduate School of Economics, Yeonsei University, 2003.

[5] Y. J. Kim, "Comparative analysis of deep-learning based english text-to-speech models," Korea University Graduate School, Feb. 2021.

[6] T.-H. Kim, "Forensic analytical study of artificial voices and modulated voices under

telecommunication channel," *Korean Digital Forensics Society, Forensic Sound Section*, pp. 28-39, Aug. 2009.

[7] S. Y. Ahn, et al., "Trends on distributed frameworks for deep learning," *2016 Electr. and Telecommun. Trends*, vol. 31, no. 3, pp. 131-141, 2016.

[8] Y. Lee, "Speech/Audio processing based on deep learning," *Broadcasting and Media Mag.*, vol. 22, no. 1, pp. 47-58, 2017.

[9] H. Y. Jung, "Speaker recognition technology - review and prospects," *KIISE*, vol. 19, no. 7, pp. 218-225, Jul. 2001.

[10] N. Dehak, P. Kenny, R. Dehak, P. Dumouchei, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 4, pp. 788-798, May 2011. (http://doi.org/10.1109/TASL.2010.2064307)

[11] Gang LV, "Joint factor analysis of channel mismatch in whispering speaker verification," Soochow University Suzhou, Nov. 2012. (http://doi.org/10.2478/v10168-012-0065-9)

[12] S. H. Mun, "DNN speaker identification algorithm based on i-vector considering speaker characteristics," *J. KICS*, vol. 66, no. 1, pp. 1073-1074, Jun. 2018.

[13] Y. Jung, "Group-based speaker embeddings for text-independent speaker verification," *The J. Acoustical Soc. Korea*, vol. 40, no. 5, pp. 496-502, Aug. 2021.

[14] Y. J. Yang, "The evolution of tele-financial fraud: an analysis of offender-victim interaction structures and respondence to 'Voice Phising'," Kyungnam University, Graduate School, 2008.

[15] H. M. Shin, "Study on telecommunications-based financial fraud based on crime script analysis and suggestions for the response plans," The Graduate School of Police Studies, Dec. 2022.

[16] S.-R. Hyun, "A study on the evolution and institutional response of voice-phishing crime methods," Korea University, 2021.

[17] KCSC(Korea Communications Standards Commission) Standards Commission. (Period : 2020.6.25 ~ 2021.9.24)

[18] S. Y. Park, "Trends and implications of phishing fraud legislation and policies in major countries," *National Assembly Research Service, NARS*, Oct. 2021.

[19] J.-K. Ji, et al., "Design and implementation of speaker verification system using voice," *J. Korean Inst. Office Automation*, vol. 5, no. 3, pp. 91-98, 2000.

김 소 운 (Sowoon Kim)

2022년 3월~현재 : 용인대학교 컴퓨터과학과 학사과정

<관심분야> 인공지능, 빅데이터, 정보보호

이 성 택 (Sungtaek Lee)

2003년 2월 : 숭실대학교 경영학부 학사

2011년 2월 : 숭실대학교 회계세무학과 석사

2018년 2월 : 숭실대학교 IT정책경영학과 박사

2020년 4월~현재 : 용인대학교 AI학부 교수

<관심분야> 기술가치평가, 기술사업화, 데이터 비즈니스모델링

[ORCID:0000-0001-9909-7698]