발명의 설명

발명의 명칭: 생성형 인공지능에 의해 생성된 딥보이스를 검출하고 관련 사고를 예방하는 인공지능 솔루션 제공 방법, 장치 및 시스템

기술분야

[1] 본 발명은 생성형 인공지능(Generative AI) 기술을 이용하여 생성된 딥보이스(Deep Voice)를 검출하고, 관련 사고를 예방하는 인공지능 솔루션 제공 방법, 장치 및 시스템에 관한 것이다.

배경기술

[2] 최근 인공지능 기술의 발전과 함께, 사람의 목소리를 매우 정교하게 모방하는 딥보이스 기술이 등장했다. 이러한 딥보이스 기술은 음성 합성, 엔터테인먼트 등 다양한 분야에서 활용될 수 있지만, 악의적인 목적으로 사용될 경우 보이스피싱, 사칭, 명예훼손 등 심각한 사회적 문제를 야기할 수 있다.

[3] 기존의 딥보이스 탐지 기술은 주로 음성 신호의 특징을 분석하는 방법을 사용한다. 예를 들어, 멜 주파수 켑스트럼(MFCC) 분석, 선형 예측 코딩(LPC) 분석 등을 통해 딥보이스에서 나타나는 특징을 추출하고, 이를 기반으로 딥보이스 여부를 판별한다. 그러나 이러한 방법들은 생성형 AI 기술의 발전으로 인해 점점 더 정교해지는 딥보이스를 효과적으로 탐지하는 데 한계를 보이고 있다.

발명의 내용

기술적 과제

[4] 본 발명은 생성형 AI 기술의 발전으로 인해 더욱 정교해진 딥보이스를 효과적으로 검출하고, 관련 사고를 예방하는 솔루션을 제공하는 것을 목표로 한다.

과제 해결 수단

[5] 본 발명의 일 형태에 따르면, 생성형 인공지능에 의해 생성된 딥보이스를 검출하는 시스템에 있어서, 인스트럭션들을 저장하는 메모리; 및 상기 인스트럭션들을 실행하는 하나 이상의 프로세서를 포함하고, 상기 프로세서는, 입력된 음성 데이터에서 주파수 특성 및 시간적 변화 패턴을 포함하는 특징 벡터를 추출하고, 상기 추출된 특징 벡터를 음향 모델인 제1 AI 모델, 언어 모델인 제2 AI 모델, 및 패턴 인식 모델인 제3 AI 모델을 포함하는 앙상블 인공지능 모델에 입력하여 각각의 개별 딥보이스 확률을 산출하며, 상기 입력된 음성 데이터의 신호 품질(Signal Quality)을 측정하여 신호 대 잡음비(SNR)를 산출하고, 상기 산출된 SNR 값에 기초하여 상기 각 AI 모델에 적용될 가중치를 실시간으로 변동시키는 적응형 가중치 할당(Adaptive Weight Allocation) 알고리즘을 통해 최종 딥보이스 확률을 연산하고, 상기 최종 딥보이스 확률이 임계값을 초과하는 경우, 딥보이스 의심

경보를 생성하도록 구성된, 딥보이스 검출 및 사고 예방 인공지능 솔루션 제공 시스템이 제공된다.

[6] 본 발명의 일 실시예에 따르면, 상기 프로세서는, 컴퓨팅 자원의 부하를 제어하기 위하여 동적 리소스 스위칭(Dynamic Resource Switching)을 수행하되,상기 입력된 음성 데이터를 연산량이 기설정된 기준 미만인 상기 제3 AI 모델을 이용하여 1차 분석하고, 상기 1차 분석 결과 딥보이스 확률이 예비 임계값을 초과하는 구간이 검출된 경우에 한하여, 상기 메모리의 활성 영역에 딥러닝 기반의 상기 제1 AI 모델 및 제2 AI 모델을 로드(Load)하여 정밀 분석을 수행하도록 구성될 수 있다.

[7] 본 발명의 일 실시예에 따르면, 상기 프로세서는, 상기 최종 딥보이스 확률이 상기 임계값을 초과하는 경우, 사용자 단말을 통해 발신자의 신원을 확인하기 위한 추가 질문을 제시하고, 상기 추가 질문에 대한 발신자의 응답 음성 신호를 수신하여, 질의 시점부터 응답 시점까지의 지연 시간(Latency) 및 응답 음성 파형의 지터(Jitter)와 쉬머(Shimmer) 변동률을 측정하며, 상기 지연 시간이 기설정된 인간 반응 범위를 벗어나거나, 상기 지터 변동률이 기설정된 인공 합성음 패턴 범위에 포함되는 경우, 상기 응답의 내용과 무관하게 상기 최종 딥보이스 확률에 페널티 가중치를 가산하여 업데이트하는 생체 활성 감지(Liveness Detection) 피드백을 수행하도록 구성될 수 있다.

[8] 본 발명의 일 실시예에 따르면, 상기 적응형 가중치 할당 알고리즘은, 상기 SNR 값이 기준치 미만인 저품질 신호일 경우, 잡음에 민감한 상기 제1 AI 모델(음향)의 가중치를 감소시키고 문맥 의존적인 상기 제2 AI 모델(언어)의 가중치를 증가시키며, 상기 최종 딥보이스 확률(P_{final})을 다음의 수식에 기초하여 산출하며, 여기서, w_i는 SNR에 따라 변동되는 각 모델의 가중치, p_i는 각 모델의 출력 확률일 수 있다.

[9]
$$P_{final} = \frac{\sum(w_i \times p_i)}{\sum w_i}$$

[10] 본 발명의 일 실시예에 따르면, 상기 제1 AI 모델은 멜 스펙트로그램(Mel-Spectrogram)을 입력받는 CNN(Convolutional Neural Network) 기반 모델이고, 상기 제2 AI 모델은 텍스트로 변환된 발화 내용을 입력받는 트랜스포머(Transformer) 기반 언어 모델이며,상기 제3 AI 모델은 음성의 피치(Pitch) 및 에너지(Energy) 변화 통계치를 입력받는 패턴 인식 모델일 수 있다.

[11] 본 발명의 다른 형태에 따르면, 컴퓨팅 장치에 의해 수행되는 딥보이스 검출 방법에 있어서, 입력된 음성 데이터에서 주파수 특성 및 시간적 변화 패턴을 포함하는 특징 벡터를 추출하는 단계; 상기 추출된 특징 벡터를 음향 모델인 제1 AI 모델, 언어 모델인 제2 AI 모델, 및 패턴 인식 모델인 제3 AI 모델을 포함하는 앙상블 모델에 입력하여 개별 확률을 산출하는 단계; 상기 입력된 음성 데이터의 신호 품질을 분석하여 신호 대 잡음비(SNR)를 측정하는 단계; 상기 측정된 SNR

값에 따라 상기 복수의 AI 모델 각각에 적용할 가중치를 동적으로 결정하는 적응형 가중치 할당을 수행하여 최종 딥보이스 확률을 연산하는 단계; 및 상기 최종 딥보이스 확률이 임계값을 초과하는 경우, 사용자 단말에 경고 메시지를 표시하거나 수사 기관 서버로 신고 신호를 전송하는 단계를 포함하는, 딥보이스 검출 및 사고 예방 인공지능 솔루션 제공 방법이 제공된다.

[12] 본 발명의 또 다른 형태에 따르면, 상기 방법을 컴퓨터에서 실행시키기 위한 프로그램을 기록한 컴퓨터 판독 가능 기록 매체가 제공된다.

발명의 효과

[13] 생성형 AI 모델을 이용하여 기존 방법보다 더욱 정확하게 딥보이스를 검출할 수 있다. 다양한 생성형 AI 모델에 대한 학습을 통해, 다양한 유형의 딥보이스를 효과적으로 탐지할 수 있다.

[14] 음성 통화, 온라인 회의 등 실시간으로 발생하는 딥보이스를 즉시 탐지하여 피해를 예방할 수 있다. 그리고 딥보이스를 이용한 보이스피싱, 사칭 등의 범죄를 예방할 수 있다.

도면의 간단한 설명

[15] 도 1은 일 실시예에 따른 생성형 인공지능에 의해 생성된 딥보이스를 검출하고 관련 사고를 예방하는 인공지능 솔루션 제공 시스템을 설명하기 위한 도면이다.

[16] 도 2은 일 실시예에 따른 뉴럴 네트워크의 학습을 설명하기 위한 도면이다.

[17] 도 3은 일 실시예에 따른 인공지능 모델의 구성을 도시한 도면이다.

[18] 도 4는 일 실시예에 따른 생성형 인공지능에 의해 생성된 딥보이스를 검출하고 관련 사고를 예방하는 인공지능 솔루션 제공 시스템의 구성을 블록도로 나타낸 것이다.

[19] 도 5는 일 실시예에 따른 생성형 인공지능에 의해 생성된 딥보이스를 검출하고 관련 사고를 예방하는 인공지능 솔루션 제공 방법의 순서도를 나타낸 것이다.

[20] 도 6은 일 실시예에 따른 생성형 인공지능에 의해 생성된 딥보이스를 검출하고 관련 사고를 예방하는 인공지능 솔루션 제공 방법의 흐름도를 나타낸 것이다.

[21] 도 7은 일 실시예에 따른 생성형 인공지능에 의해 생성된 딥보이스를 검출하고 관련 사고를 예방하는 인공지능 솔루션 제공 방법의 흐름도를 나타낸 것이다.

[22] 도 8은 본 발명의 일 실시예에 따른 적응형 가중치 할당 프로세스를 도시한다.

발명의 실시를 위한 형태

[23] 이하에서, 첨부된 도면을 참조하여 실시예들을 상세하게 설명한다.

[24] 그러나, 실시예들에는 다양한 변경이 가해질 수 있어서 특허출원의 권리 범위가 이러한 실시예들에 의해 제한되거나 한정되는 것은 아니다. 실시예들에 대한 모든 변경, 균등물 내지 대체물이 권리 범위에 포함되는 것으로 이해되어야 한다.

[25] 실시예들에 대한 특정한 구조적 또는 기능적 설명들은 단지 예시를 위한 목적으로 개시된 것으로서, 다양한 형태로 변경되어 실시될 수 있다. 따라서, 실시예

들은 특정한 개시형태로 한정되는 것이 아니며, 본 명세서의 범위는 기술적 사상에 포함되는 변경, 균등물, 또는 대체물을 포함한다.

[26] 제1 또는 제2 등의 용어를 다양한 구성요소들을 설명하는데 사용될 수 있지만, 이런 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 해석되어야 한다. 예를 들어, 제1 구성요소는 제2 구성요소로 명명될 수 있고, 유사하게 제2 구성요소는 제1 구성요소로도 명명될 수 있다.

[27] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다.

[28] 실시예에서 사용한 용어는 단지 설명을 목적으로 사용된 것으로, 한정하려는 의도로 해석되어서는 안 된다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 명세서 상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[29] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 실시예가 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가지고 있다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥 상 가지는 의미와 일치하는 의미를 가지는 것으로 해석되어야 하며, 본 출원에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[30] 또한, 첨부 도면을 참조하여 설명함에 있어, 도면 부호에 관계없이 동일한 구성요소는 동일한 참조부호를 부여하고 이에 대한 중복되는 설명은 생략하기로 한다. 실시예를 설명함에 있어서 관련된 공지 기술에 대한 구체적인 설명이 실시예의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우 그 상세한 설명을 생략한다.

[31] 실시예들은 퍼스널 컴퓨터, 랩톱 컴퓨터, 태블릿 컴퓨터, 스마트폰, 텔레비전, 스마트 가전 기기, 지능형 자동차, 키오스크, 웨어러블 장치 등 다양한 형태의 제품으로 구현될 수 있다.

[32] 인공지능(Artificial Intelligence, AI) 시스템은 인간 수준의 지능을 구현하는 컴퓨터 시스템이며, 기존 규칙(Rule) 기반의 스마트 시스템과 달리 기계가 스스로 학습하고 판단하는 시스템이다. 인공지능 시스템은 사용할수록 인식률이 향상되고 사용자 취향을 보다 정확하게 이해할 수 있게 되어, 기존 규칙 기반의 스마트 시스템은 점차 심층 학습(Deep Learning) 기반 인공지능 시스템으로 대체되고 있다.

[33] 인공지능 기술은 기계 학습 및 기계 학습을 활용한 요소기술들로 구성된다. 기계 학습은 입력 데이터들의 특징을 스스로 분류/학습하는 알고리즘 기술이며,

요소기술은 심층 학습 등의 기계 학습 알고리즘을 활용하여 인간두뇌의 인지, 판단 등의 기능을 모사하는 기술로서, 언어적 이해, 시각적 이해, 추론/예측, 지식 표현, 동작 제어 등의 기술 분야로 구성된다.

[34] 인공지능 기술이 응용되는 다양한 분야는 다음과 같다. 언어적 이해는 인간의 언어/문자를 인식하고 응용/처리하는 기술로서, 자연어 처리, 기계 번역, 대화시스템, 질의 응답, 음성 인식/합성 등을 포함한다. 시각적 이해는 사물을 인간의 시각처럼 인식하여 처리하는 기술로서, 객체 인식, 객체 추적, 영상 검색, 사람 인식, 장면이해, 공간 이해, 영상 개선 등을 포함한다. 추론 예측은 정보를 판단하여 논리적으로 추론하고 예측하는 기술로서, 지식/확률 기반 추론, 최적화 예측, 선호 기반 계획, 추천 등을 포함한다. 지식 표현은 인간의 경험정보를 지식데이터로 자동화 처리하는 기술로서, 지식 구축(데이터 생성/분류), 지식 관리(데이터 활용) 등을 포함한다. 동작 제어는 차량의 자율 주행, 로봇의 움직임을 제어하는 기술로서, 움직임 제어(항법, 충돌, 주행), 조작 제어(행동 제어) 등을 포함한다.

[35] 일반적으로 기계 학습 알고리즘을 실생활에 적용하기 위해서는 기계학습의 기본 방법론의 특성상 Trial and Error 방식으로 학습을 수행하게 된다. 특히, 심층 학습의 경우 수십만 번의 반복 실행을 필요로 한다. 이를 실제 물리적인 외부 환경에서 실행하기는 불가능하여 대신 실제 물리적인 외부 환경을 컴퓨터상에서 가상으로 구현하여 시뮬레이션을 통해 학습을 수행한다.

[36] 본 발명에서, 인공지능(Artificial Intelligence, AI)은 인간의 학습능력, 추론능력, 지각능력 등을 모방하고, 이를 컴퓨터로 구현하는 기술을 의미하고, 기계 학습, 심볼릭 로직(Symbolic Logic) 등의 개념을 포함할 수 있다. 기계학습(Machine Learning, ML)은 입력 데이터들의 특징을 스스로 분류 또는 학습하는 알고리즘 기술이다. 인공지능의 기술은 기계 학습의 알고리즘으로써 입력 데이터를

[37] 분석하고, 그 분석의 결과를 학습하며, 그 학습의 결과에 기초하여 판단이나 예측을 할 수 있다. 또한, 기계 학습의 알고리즘을 활용하여 인간 두뇌의 인지, 판단 등의 기능을 모사하는 기술들 역시 인공지능의 범주로 이해될 수 있다. 예를 들어, 언어적 이해, 시각적 이해, 추론/예측, 지식 표현, 동작 제어 등의 기술 분야가 포함될 수 있다.

[38] 인공지능 학습모델 또는 신경망 모델은 인간의 뇌 구조를 컴퓨터 상에서 구현하도록 설계될 수 있으며, 인간의 신경망의 뉴런(neuron)을 모의하며 가중치를 가지는 복수의 네트워크 노드들을 포함할 수 있다. 복수의 네트워크 노드들은 뉴런이 시냅스(synapse)를 통하여 신호를 주고받는 뉴런의 시냅틱(synaptic) 활동을 모의하여, 서로 간의 연결 관계를 가질 수 있다. 인공지능 학습모델에서 복수의 네트워크 노드들은 서로 다른 깊이의 레이어에 위치하면서 컨볼루션(convolution) 연결 관계에 따라 데이터를 주고받을 수 있다. 인공지능 학습모델은, 예를 들어, 인공 신경망 모델(Artificial Neural Network), 컨볼루션 신경망 모델(Convolution Neural Network: CNN) 등일 수 있다. 일 실시예로서, 인공지능 학습모델은, 지도학

습(Supervised Learning), 비지도 학습(Unsupervised Learning), 강화 학습(Reinforcement Learning) 등의 방식에 따라 기계 학습될 수 있다. 기계 학습을 수행하기 위한 기계 학습 알고리즘에는, 의사결정트리(Decision Tree), 베이지안망(Bayesian Network), 서포트 벡터 머신(Support Vector Machine), 인공 신경망(Artificial Neural Network), 에이다부스트(Ada-boost), 퍼셉트론(Perceptron), 유전자 프로그래밍(Genetic Programming), 군집화(Clustering) 등이 사용될 수 있다.

[39] 이중, CNN은 최소한의 전처리(preprocess)를 사용하도록 설계된 다계층 퍼셉트론(multilayer perceptrons)의 한 종류이다. CNN은 하나 또는 여러 개의 합성곱 계층과 그 위에 올려진 일반적인 인공 신경망 계층들로 이루어져 있으며, 가중치와 통합 계층(pooling layer)들을 추가로 활용한다. 이러한 구조 덕분에 CNN은 2차원 구조의 입력 데이터를 충분히 활용할 수 있다. 다른 딥러닝 구조들과 비교해서, CNN은 영상, 음성 분야 모두에서 좋은 성능을 보여준다. CNN은 또한 표준 역전달을 통해 훈련될 수 있다. CNN은 다른 피드포워드 인공신경망 기법들보다 쉽게 훈련되는 편이고 적은 수의 매개변수를 사용한다는 이점이 있다.

[40] 컨볼루션 네트워크는 묶인 파라미터들을 가지는 노드들의 집합들을 포함하는 신경 네트워크들이다. 사용 가능한 트레이닝 데이터의 크기 증가와 연산 능력의 가용성이, 구분적 선형 단위 및 드롭아웃 트레이닝과 같은 알고리즘 발전과 결합되어, 많은 컴퓨터 비전 작업들이 크게 개선되었다. 오늘날 많은 작업에 사용할 수 있는 데이터 세트들과 같은 엄청난 양의 데이터 세트에서는 초과 맞춤(outfitting)이 중요하지 않으며, 네트워크의 크기를 늘리면 테스트 정확도가 향상된다. 컴퓨팅 리소스들의 최적 사용은 제한 요소가 된다. 이를 위해, 심층 신 경 네트워크들의 분산된, 확장 가능한 구현예가 사용될 수 있다.

[41] 도 1은 일 실시예에 따른 생성형 인공지능에 의해 생성된 딥보이스를 검출하고 관련 사고를 예방하는 인공지능 솔루션 제공 시스템을 설명하기 위한 도면이다.

[42] 도 1에 도시한 바와 같이, 생성형 인공지능에 의해 생성된 딥보이스를 검출하고 관련 사고를 예방하는 인공지능 솔루션 제공 시스템(100)은, 다수의 사용자 단말(110-1, ... 110-n), 서버(120) 및 데이터베이스(130)를 포함할 수 있다. 일 실시예에 따르면, 데이터베이스(130)는 서버(120)와 별도로 구성된 것으로 도시되었지만 이에 한정되지 않고, 데이터베이스(130)가 서버(120)내에 구비될 수도 있다. 예를 들어, 서버(120)는, 기계 학습 알고리즘의 수행을 위한 다수의 인공지능을 포함할 수 있다. 일 실시예에 따르면, 다수의 사용자 단말(110-1, ... 110-n), 서버(120) 및 데이터베이스(130)는 네트워크(N)를 통하여 서로 통신 가능하도록 연결될 수 있다.

[43] 네트워크(N)는, 다수의 사용자 단말(110-1, ... 110-n, 서버(120), 데이터베이스(130) 등 간의 무선 또는 유선 통신을 수행하도록 할 수 있다. 예를 들어, 네트워크는 LTE(long-term evolution), LTE-A(LTE Advanced), CDMA(code division multiple access), WCDMA(wideband CDMA), WiBro(Wireless BroadBand), WiFi(wireless fidelity), 블루투스(Bluetooth), NFC(near field communication), GPS(Global Position

ing System) 또는 GNSS(global navigation satellite system) 등의 방식에 따른 무선 통신을 수행하도록 할 수 있다. 예를 들어, 네트워크 (N)는 USB(universal serial bus), HDMI(high definition multimedia interface), RS 232(recommended standard 232) 또는 POTS(plain old telephone service) 등의 방식에 따른 유선 통신을 수행하도록 할 수도 있다.

[44] 데이터베이스(130)는, 다양한 데이터를 저장할 수 있다. 데이터베이스(130)에 저장되는 데이터는, 다수의 사용자 단말(110-1, ... 110-n), 서버(120)의 적어도 하나의 구성요소에 의해 획득되거나, 처리되거나, 사용되는 데이터로서, 소프트웨어(예를 들어: 프로그램)를 포함할 수 있다. 데이터베이스(130)는, 휘발성 및/또는 비휘발성 메모리를 포함할 수 있다.

[45] 도 2은 일 실시예에 따른 뉴럴 네트워크의 학습을 설명하기 위한 도면이다.

[46] 도 2에 도시한 바와 같이, 학습 장치는 다수의 사용자 단말(110-1, ... 110-n)로부터 수신된 리뷰 응답을 항목별로 하기 위하여 뉴럴 네트워크(123)를 학습시킬 수 있다. 또한, 학습 장치는 사용자의 이동 경로 정보로부터 사용자 체류 내역을 추출하기 위하여 뉴럴 네트워크(123)를 학습시킬 수 있다. 일 실시예에 따르면, 학습 장치는 서버 (120)와 다른 별개의 주체일 수 있지만, 이에 제한되는 것은 아니다.

[47] 뉴럴 네트워크(123)는 트레이닝 샘플들이 입력되는 입력 레이어(121)와 트레이닝 출력들을 출력하는 출력 레이어(125)를 포함하고, 트레이닝 출력들과 레이블들 사이의 차이에 기초하여 학습될 수 있다. 여기서, 레이블들은 리뷰 응답에 대응하는 항목들에 기초하여 정의되고, 이동 경로 정보에 대응하는 사용자 체류 내역에 기초하 여 정의될 수 있다. 뉴럴 네트워크(123)는 복수의 노드들의 그룹으로 연결되어 있고, 연결된 노드들 사이의 가중치들과 노드들을 활성화시키는 활성화 함수에 의해 정의된다.

[48] 학습 장치는 GD(Gradient Decent) 기법 또는 SGD(Stochastic Gradient Descent) 기법을 이용하여 뉴럴 네트워크(123)를 학습시킬 수 있다. 학습 장치는 뉴럴 네트워크의 출력들 및 레이블들 의해 설계된 손실 함수(Loss Function)를 이용할 수 있다.

[49] 학습 장치는 미리 정의된 손실 함수를 이용하여 트레이닝 에러를 계산할 수 있다. 손실 함수는 레이블, 출력 및 파라미터를 입력 변수로 미리 정의될 수 있고, 여기서 파라미터는 뉴럴 네트워크(123) 내 가중치들에 의해 설정될 수 있다. 예를 들어, 손실 함수는 MSE(Mean Square Error) 형태, 엔트로피(entropy) 형태 등으로 설계될 수 있는데, 손실 함수가 설계되는 실시예에는 다양한 기법 또는 방식이 채용될 수 있다.

[50] 학습 장치는 역전파(Backpropagation) 기법을 이용하여 트레이닝 에러에 영향을 주는 가중치들을 찾아낼 수 있다. 여기서, 가중치들은 뉴럴 네트워크(123) 내 노드들 사이의 관계들이다. 학습 장치는 역전파 기법을 통해 찾아낸 가중치들을 최적화시키기 위해 레이블들 및 출력들을 이용한 SGD 기법을 이용할 수 있다.

예를 들어, 학습 장치는 레이블들, 출력들 및 가중치들에 기초하여 정의된 손실 함수의 가중치들을 SGD 기법을 이용하여 갱 신할 수 있다.

[51] 일 실시예에 따르면, 학습 장치는 리뷰 응답으로부터 제1 객체들을 추출하며, 제1 객체들에 대응하는 항목들인 제1 레이블들을 획득하고, 제1 객체들을 제1 뉴럴 네트워크로 적용하여, 제1 객체들에 대응하는 제1 트레이닝 출력들을 생성하며, 제1 트레이닝 출력들, 제1 레이블들에 기초하여, 제1 뉴럴 네트워크를 학습시킬 수 있다.

[52] 학습 장치는 이동 경로 정보로부터 제2 객체들을 추출하며, 제2 객체들에 대응하는 사용자 체류 내역인 제2 레이블들을 획득하고, 제2 객체들을 제2 뉴럴 네트워크로 적용하여, 제2 객체들에 대응하는 제2 트레이닝 출력들을 생성하며, 제2 트레이닝 출력들, 제2 레이블들에 기초하여, 제2 뉴럴 네트워크를 학습시킬 수 있다.

[53] 일 실시예에 따르면, 학습 장치는 리뷰 응답의 구성 특징들, 위치 특징들 및 패턴 특징들에 기초하여 제1 트레이닝 특징 벡터들을 생성할 수 있다.

[54] 특징을 추출하는 데는 다양한 방식이 채용될 수 있다.

[55] 일 실시예에 따르면, 학습 장치는 이동 경로 정보의 구성 특징들, 길이 특징들 및 패턴 특징들에 기초하여 제2 트레이닝 특징 벡터들을 생성할 수 있다. 특징을 추출하는 데는 다양한 방식이 채용될 수 있다.

[56] 일 실시예에 따르면, 학습 장치는 제1 트레이닝 특징 벡터들을 뉴럴 네트워크(123)에 적용하여 트레이닝 출력들을 획득할 수 있다. 학습 장치는 트레이닝 출력들과 제1 레이블들에 기초하여 뉴럴 네트워크(123)의 리뷰 항목 추출 알고리즘을 학습시킬 수 있다. 학습 장치는 트레이닝 출력들에 대응하는 트레이닝 에러들을 계산하고, 그 트레이닝 에러들을 최소화하기 위해 뉴럴 네트워크(123) 내 노드들의 연결 관계를 최적화하여 뉴럴 네트워크(123)의 리뷰 항목 추출 알고리즘을 학습시킬 수 있다. 서버(120)는 학습이 완료된 제1 뉴럴 네트워크를 이용하여 리뷰 응답으로부터 항목들을 추출할 수 있다.

[57] 일 실시예에 따르면, 학습 장치는 제2 트레이닝 특징 벡터들을 뉴럴 네트워크(123)에 적용하여 트레이닝 출력들을 획득할 수 있다. 학습 장치는 트레이닝 출력들과 제2 레이블들에 기초하여 뉴럴 네트워크(123)의 사용자 체 류 내역 획득 알고리즘을 학습시킬 수 있다. 학습 장치는 트레이닝 출력들에 대응하는 트레이닝 에러들을 계산하고, 그 트레이닝 에러들을 최소화하기 위해 뉴럴 네트워크(123) 내 노드들의 연결 관계를 최적화하여 뉴럴 네트워크(123)의 사용자 체류 내역 획득 알고리즘을 학습시킬 수 있다. 서버(120)는 학습이 완료된 제2 뉴럴 네트워크를 이용하여 이동 경로 정보부터 사용자 체류 내역을 획득할 수 있다.

[58] 도 3은 일 실시예에 따른 인공지능 모델의 구성을 도시한 도면이다.

[59] 일 실시예에 따른 인공지능 모델은 입력 계층(input layer), 은닉 계층(hidden layer) 및 출력 계층(output layer)을 포함할 수 있다.

[60] 입력 계층(input layer)은 인공지능 모델에 입력되는 입력 값과 관련된 계층이다.

[61] 은닉 계층(hidden layer)에서는 입력 값에 대하여 MAC 연산(multiply-accumulate)과 활성화 연산을 수행하여 피쳐 맵을 출력할 수 있다.

[62] MAC 연산은 입력 값과 대응되는 가중치를 각각 곱하고, 곱한 값들을 합하는 연산일 수 있다.

[63] 활성화 연산은 MAC 연산의 결과를 활성화 함수에 입력하여 결과 값을 출력하는 연산일 수 있다. 활성화 함수는, 다양한 유형일 수 있다. 예를 들어, 활성화 함수는, 시그모이드 함수, 탄젠트 함수, 렐루 함수, 리키 렐루 함수, 맥스 아웃 함수 및/또는 엘루 함수을 포함할 수 있으나 그 종류에 제한이 없다.

[64] 은닉 계층은 적어도 하나의 계층(layer)로 구성될 수 있다. 예를 들어, 은닉 계층이 제 1 은닉 계층 및 제 2 은닉 계층으로 구성된 경우, 제 1 은닉 계층은 입력 계의 입력 값에 기반하여 MAC 연산 및 활성화 연산을 수행하여 피쳐 맵을 출력하고, 제 1 은닉 계층에서의 결과 값인 피쳐 맵이 제 2 은닉 계층에서의 입력 값이 될 수 있다. 제 2 은닉 계층은 제 1 은닉 계층의 결과 값인 피쳐 맵에 기반하여 MAC 연산 및 활성화 연산을 수행할 수 있다.

[65] 출력 계층(output layer)은, 은닉 계층에서 수행한 연산의 결과 값과 관련된 계층일 수 있다.

[66] 일 실시예에서, 학습 모델은 주어진 말뭉치에서 빈번히 결합하여 사용된 음절(글자) 패턴을 학습 하여 복합어 및 개체명의 경계를 자동으로 학습하고, 제1 UI 소스의 객체 정보와 브라우저에서 렌더링한 객체 정보를 통합하여 학습용 객체 정보 파일을 생성하고, 상기 학습용 객체 정보 파일을 이용하여 딥러닝 네트워크의 학습을 위한 학습 데이터를 생성하고, 지원 시스템의 다양한 도메인들의 데이터를 수신하고, 상기 다양한 도메인들 각각에 대응하는 적어도 하나의 표준화 방법에 기반하여, 상기 다양한 도메인들의 데이터를 통합된 형식으로 표준화하며, 특정 도메인의 데이터를 학습 및 추론하고, 상기 특정 도메인의 데이터에서, 상기 표준화를 위해 전달될 정보를 결정하고, 상기 다양한 도메인들로부터의 데이터에 대해 후처리(post processing)를 수행할 수 있다. 제1 UI 소스는 XML 파일을 포함하고, 상기 학습용 객체 정보 파일은 특징(Features) 학습을 위한 입력 JSON 파일과 학습시 정답(Label) 데이터인 출력 JSON 파일을 포함하며, 상기 출력 JSON 파일은 웹표준을 준수하여 구현된 HTML의 DOM Tree 정보가 포함된 파일을 포함하고 상기 다양한 도메인들은 RAN(radio access network), 트랜스포트(transport) 또는 코어(core) 중 적어도 하나를 포함하고, 후처리는 상관(correlation) 기능을 포함할 수 있다.

[67] 도 4는 일 실시예에 따른 생성형 인공지능에 의해 생성된 딥보이스를 검출하고 관련 사고를 예방하는 인공지능 솔루션 제공 시스템의 구성을 블록도로 나타낸 것이다.

[68] 일 실시예에 따른 시스템(400)은 프로세서(420) 및 메모리(430)를 포함할 수 있으며, 도시된 구성 중 일부가 생략 또는 치환될 수도 있다. 일 실시예에 따른 시스템(400)는 서버 또는 단말일 수 있다. 일 실시예에 따르면, 프로세서(420)는 시스

템(400)의 각 구성 요소들의 제어 및/또는 통신에 관한 연산이나 데이터 처리를 수행할 수 있는 구성으로써, 하나 이상의 프로세서들로 구성될 수 있다. 메모리(430)는 상술한 방법과 관련된 정보를 저장하거나 상술한 방법이 구현된 프로그램을 저장할 수 있다. 메모리(430)는 휘발성 메모리 또는 비휘발성 메모리일 수 있다. 메모리(430)는 다양한 파일 데이터들을 저장할 수 있으며, 프로세서(120)의 동작에 따라 저장된 파일 데이터들은 업데이트 될 수 있다.

[69] 일 실시예에 따르면, 프로세서(420)는 프로그램을 실행하고, 장치(400)를 제어할 수 있다. 프로세서(120)에 의하여 실행되는 프로그램의 코드는 메모리(430)에 저장될 수 있다. 프로세서(420)의 동작들은 메모리(430)에 저장된 인스트럭션들을 로딩(loading)함으로써 수행될 수 있다. 시스템(400)은 입출력 장치(도면 미 표시)를 통하여 외부 장치(예를 들어, 퍼스널 컴퓨터 또는 네트워크)에 연결되고, 데이터를 교환할 수 있다.

[70] 일 실시예에 따른 시스템(400)은 프로세서(420) 및 메모리(430)를 포함한다. 일 실시예에 따른 시스템(400)은 상술한 서버 또는 단말일 수 있다. 프로세서(420)는 도 1 내지 도 3을 통하여 전술한 적어도 하나의 장치들을 포함하거나, 도 1 내지 도 3을 통하여 전술한 적어도 하나의 방법을 수행할 수 있다. 메모리(430)는 상술한 방법과 관련된 정보를 저장하거나 상술한 방법이 구현된 프로그램을 저장할 수 있다. 메모리(430)는 휘발성 메모리 또는 비휘발성 메모리일 수 있다.

[71] 일 실시예에 따른 시스템(400)은 프로세서(420) 및 메모리(430)를 포함한다. 일 실시예에 따른 시스템(400)은 상술한 서버 또는 단말일 수 있다. 프로세서(420)는 도 1 내지 도 3을 통하여 전술한 적어도 하나의 장치들을 포함하거나, 도 1 내지 도 3을 통하여 전술한 적어도 하나의 방법을 수행할 수 있다. 메모리(430)는 상술한 방법과 관련된 정보를 저장하거나 상술한 방법이 구현된 프로그램을 저장할 수 있다. 메모리(430)는 휘발성 메모리 또는 비휘발성 메모리일 수 있다.

[72]

[73] 최종 딥보이스 확률(P_{final})은 제1 AI 모델(음향), 제2 AI 모델(언어), 제3 AI 모델(패턴)의 각 출력 확률(p_1, p_2, p_3)에 가중치(w_1, w_2, w_3)를 적용하여 산출된다.

[74]

$$P_{final} = \frac{w_1 \cdot p_1 + w_2 \cdot p_2 + w_3 \cdot p_3}{w_1 + w_2 + w_3}$$

[75] 이때, 각 가중치 w_i는 고정된 값이 아니라, 입력된 음성 신호의 신호 대 잡음비(SNR) 및 각 모델의 신뢰도 점수(Confidence Score)에 기반하여 실시간으로 변동된다. 예를 들어, 입력 음성의 배경 잡음이 심하여 SNR이 기준치(예: 10dB) 미만인 경우, 잡음에 취약한 제1 AI 모델(음향)의 가중치 w_1은 0.5배로 감소시키고, 문맥 의존적인 제2 AI 모델(언어)의 가중치 w_2를 1.5배로 증가시키는 적응형 가중치 할당(Adaptive Weight Allocation) 알고리즘이 적용된다.

[76]

[77] 도 5는 일 실시예에 따른 생성형 인공지능에 의해 생성된 딥보이스를 검출하고 관련 사고를 예방하는 인공지능 솔루션 제공 방법의 순서도를 나타낸 것이다.

[78] 도 5의 순서도에서 프로세스 단계들, 방법 단계들, 알고리즘들 등이 순차적인 순서로 설명되었지만, 그러한 프로세스들, 방법들 및 알고리즘들은 임의의 적합한 순서로 작동하도록 구성될 수 있다. 다시 말하면, 본 발명의 다양한 실시예들에서 설명되는 프로세스들, 방법들 및 알고리즘들의 단계들이 본 발명에서 기술된 순서로 수행될 필요는 없다. 또한, 일부 단계들이 비동시적으로 수행되는 것으로서 설명되더라도, 다른 실시예에서는 이러 한 일부 단계들이 동시에 수행될 수 있다. 또한, 도면에서의 묘사에 의한 프로세스의 예시는 예시된 프로세스 가 그에 대한 다른 변화들 및 수정들을 제외하는 것을 의미하지 않으며, 예시된 프로세스 또는 그의 단계들 중 임의의 것이 본 발명의 다양한 실시예들 중 하나 이상에 필수적임을 의미하지 않으며, 예시된 프로세스가 바람 직하다는 것을 의미하지 않는다.

[79] 동작 S510에서, 생성형 인공지능(AI)을 통해 만들어진 딥보이스를 탐지하고 관련 범죄를 예방하는 시스템(예: 도 4의 시스템(400))은 프로세서(예: 도 4의 프로세서(420))의 제어 하에, 딥보이스 의심 특징을 추출할 수 있다. 시스템(400)은 음성 데이터(예: 전화 통화)를 실시간으로 입력받고, 입력된 음성 데이터에서 주파수 특성, 시간적 변화 패턴 및 미세 잡음을 분석하여 딥보이스 의심 특징을 추출할 수 있다.

[80] 예를 들어, 전화 통화 내용이 시스템에 입력되면, 시스템은 이 음성 데이터를 실시간으로 수집한다. 입력된 음성 데이터에서 주파수 특성을 분석하여 특정 주파수 대역에서 비정상적인 패턴이 발견되면, 이는 딥보이스 의심 특징이 될 수 있다. 예를 들어, 사용자가 "안녕하세요, 저는 김철수입니다"라고 말할 때, 시스템은 이 음성의 주파수 패턴을 분석하여 정상적인 화자의 음성과 다른 점을 식별한다. 또한, 시간적 변화 패턴을 분석하여 음성의 발음 속도나 억양의 변화를 감지한다. 만약 사용자가 평소보다 빠르게 말하거나 비정상적인 억양을 사용하면, 이는 딥보이스 의심 특징으로 간주된다. 마지막으로, 미세 잡음을 분석하여 배경 소음이나 음성의 왜곡 정도를 평가한다. 예를 들어, 통화 중에 발생한 기계적 소음이나 왜곡된 음성이 감지되면, 이는 추가적인 의심 요소로 작용한다.

[81] 사용자 A가 보이스피싱 범죄자 B와 통화 중이라고 가정할 수 있다. 이 경우 범죄자 B는 딥보이스 기술을 이용하여 여성의 목소리로 변조된 음성을 사용할 수 있다. 시스템(400)은 실시간으로 A와 B의 통화 내용을 입력받는다. 시스템(400)은 B의 음성에서 일반적인 여성 음성보다 낮은 주파수 대역의 에너지가 높게 나타나는 것을 감지할 수 있다. 시스템(400)은 B의 음성에서 단어 사이의 간격이나 억양 변화가 부자연스럽거나 일정한 패턴을 보이는 것을 감지한다. 시스템(400)은 B의 음성에서 딥러닝 모델 학습 데이터에 존재하는 특정 잡음 패턴이나, 사람 목소리에서는 일반적으로 나타나지 않는 미세한 디지털 잡음을 검출할 수 있다.

[82] 동작 S520에서, 시스템(400)은 딥보이스 여부를 판별할 수 있다. 시스템(400)은 생성형 AI 모델을 이용하여 추출된 특징을 분석하여 딥보이스 여부를 판별할 수 있다.

[83] AI 모델은 딥보이스 탐지에 특화된 모델로, 대량의 딥보이스 음성 데이터와 실제 사람 음성 데이터를 학습하여 딥보이스 여부를 판별하는 데 높은 정확도를 갖는다. AI 모델은 입력된 특징들을 분석한 결과, B의 음성에서 딥보이스 생성 과정에서 발생하는 특징적인 패턴들을 발견하고 딥보이스로 판별한다.

[84] 일 실시예에 따르면, 시스템(400)은 딥보이스 여부를 판별한다. 이 단계에서는 생성형 AI 모델을 활용하여 이전 단계에서 추출된 특징을 분석한다. 예를 들어, 시스템은 딥러닝 기반의 음향 모델을 사용하여 음성의 주파수 패턴과 시간적 변화 패턴을 비교 분석한다. 만약 시스템이 "이 음성이 생성형 AI에 의해 생성된 것일 가능성이 높다"라는 결론에 도달하면, 이는 딥보이스로 판별된다. 예를 들어, 특정 주파수 대역에서 비정상적인 패턴이 반복적으로 나타나거나, 발음의 일관성이 결여된 경우 시스템은 이를 딥보이스로 인식할 수 있다. 이러한 판별 과정은 실시간으로 이루어지며, 시스템은 각 모델의 출력 결과를 종합적으로 분석하여 최종 결정을 내린다.

[85] 동작 S530에서, 시스템(400)은 경고 메시지를 사용자 단말에 표시하고 자동으로 신고할 수 있다. 시스템(400)은 딥보이스로 판별될 경우, "주의! 딥보이스 의심 음성이 감지되었습니다." 라는 경고 메시지를 사용자 단말에 표시하며, 딥보이스 의심 음성 데이터와 관련된 정보를 경찰청 사이버 안전국에 자동으로 신고하도록 제어할 수 있다.

[86] 일 실시예에 따르면, 시스템(400)은 생성형 인공지능 모델(AI)을 이용하여 추출된 특징을 분석하고, 분석 결과를 바탕으로 딥보이스 확률을 계산하며, 딥보이스 확률이 지정된 수준을 초과하여 딥보이스로 판별될 경우 경고 메시지를 표시할 수 있다.

[87] 시스템은 딥보이스로 판별될 경우 경고 메시지를 사용자 단말에 표시하고 자동으로 신고하여 범죄를 예방할 수 있다. 예를 들어, 시스템이 "주의! 딥보이스 의심 음성이 감지되었습니다."라는 경고 메시지를 사용자 A의 스마트폰 화면에 표시한다. 이 메시지는 음성이 딥보이스로 판별되었음을 알리며, 사용자가 즉각적으로 상황을 인식할 수 있도록 돕는다. 또한, 시스템은 딥보이스 의심 음성 데이터와 관련된 정보를 경찰청 사이버 안전국에 자동으로 신고하도록 제어한다. 이때, 신고되는 정보에는 발신자의 전화번호, 통화 시간, 딥보이스 확률, 그리고 통화 내용의 일부가 포함된다. 예를 들어, 시스템은 "사용자 A의 전화번호 010-1234-5678에서 2024년 10월 17일 15시 30분에 발생한 딥보이스 의심 음성"이라는 형태로 신고를 진행한다.

[88] AI 모델이 B의 음성을 딥보이스로 판별하면, 시스템(400)은 즉시 A의 스마트폰 화면에 "주의! 딥보이스 의심 음성이 감지되었습니다."라는 경고 메시지를 표시한다. 동시에, B의 딥보이스 음성 데이터, 통화 시작 시간, 통화 내용 등 관련 정

보를 경찰청 사이버 안전국에 자동으로 신고한다. 이를 통해 A는 보이스피싱 범죄에 대한 경각심을 갖고 피해를 예방할 수 있으며, 경찰은 신고된 정보를 바탕으로 범죄자를 추적하고 수사에 활용할 수 있다.

[89] 도 6은 일 실시예에 따른 생성형 인공지능에 의해 생성된 딥보이스를 검출하고 관련 사고를 예방하는 인공지능 솔루션 제공 방법의 흐름도를 나타낸것이다.

[90] 도 6의 순서도에서 프로세스 단계들, 방법 단계들, 알고리즘들 등이 순차적인 순서로 설명되었지만, 그러한 프로세스들, 방법들 및 알고리즘들은 임의의 적합한 순서로 작동하도록 구성될 수 있다. 다시 말하면, 본 발명의 다양한 실시예들에서 설명되는 프로세스들, 방법들 및 알고리즘들의 단계들이 본 발명에서 기술된 순서로 수행될 필요는 없다. 또한, 일부 단계들이 비동시적으로 수행되는 것으로서 설명되더라도, 다른 실시예에서는 이러 한 일부 단계들이 동시에 수행될 수 있다. 또한, 도면에서의 묘사에 의한 프로세스의 예시는 예시된 프로세스 가 그에 대한 다른 변화들 및 수정들을 제외하는 것을 의미하지 않으며, 예시된 프로세스 또는 그의 단계들 중 임의의 것이 본 발명의 다양한 실시예들 중 하나 이상에 필수적임을 의미하지 않으며, 예시된 프로세스가 바람 직하다는 것을 의미하지 않는다.

[91] 동작 S610에서, 시스템(예: 도 4의 시스템(400))은 프로세서(예: 도 4의 프로세서(420))의 제어 하에, 입력된 음성 데이터에서 주파수 특성을 추출할 수 있다. 시스템(400)은 입력된 음성 데이터에서 멜 주파수 켑스트럼 계수(MFCC), 선형 예측 코딩(LPC) 계수, 스펙트로그램을 포함하는 주파수 특성을 추출하고, 피치 변화, 포먼트 변화 및 에너지 변화를 포함하는 시간적 변화 패턴을 추출할 수 있다.

[92] 먼저, 시스템(400)은 음성 데이터를 작은 시간 단위(예: 20ms)로 나누고 각 단위에 대해 푸리에 변환을 수행하여 주파수 스펙트럼을 얻는다. 이 스펙트럼에서 멜 주파수 켑스트럼 계수(MFCC)를 추출하는데, 이는 사람의 청각 시스템이 주파수를 인지하는 방식을 모방하여 주파수 정보를 압축한 것이다. MFCC는 음성 인식에서 널리 사용되는 특징으로, 특히 잡음에 강건하다는 장점을 가진다.

[93] 시스템(400)은 이후 선형 예측 코딩(LPC) 계수를 추출한다. LPC는 음성 신호의 스펙트럼 포락선을 효율적으로 표현하는 기술로, 현재 음성 샘플을 과거 샘플들의 선형 조합으로 예측하여 그 차이를 분석한다. 이를 통해 성도의 공명 특성을 파악하고, 음성의 특징을 효과적으로 나타낼 수 있다.

[94] 스펙트로그램은 시간에 따른 주파수 변화를 시각적으로 표현한 것으로, 음성 신호의 에너지 분포를 분석하는 데 유용하다. 프로세서(420)는 스펙트로그램을 분석하여 피치 변화, 즉 음성의 높낮이 변화를 추적한다. 또한, 포먼트 변화를 분석하는데, 포먼트는 성도의 공명 주파수로, 모음의 특징을 결정하는 중요한 요소다. 마지막으로, 시간에 따른 에너지 변화를 분석하여 음성의 강도 변화를 파악한다.

[95] 동작 S620에서, 시스템(400)은 각 모델의 출력 결과를 가중 평균하여 최종 딥보이스 확률을 계산할 수 있다. 시스템(400)은 배경 소음, 호흡 소리 및 입술 소리를

포함하는 미세 잡음을 추출하고, 음소 지속 시간, 공명 특징 및 조음 방식을 포함하는 발음 패턴을 추출하며, 단어 사이의 간격 및 문장 사이의 간격을 포함하는 음성 휴지 시간을 추출하고, 발화 내용, 대화 주제 및 상황 정보를 포함하는 문맥 정보를 추출하며, 상기 추출된 특징들을 딥러닝 기반의 음향 모델, 자연어 처리 기반의 언어 모델, 및 통계 기반의 패턴 인식 모델을 포함하는 앙상블 AI 모델에 입력하여 딥보이스 여부를 판별하고, 각 모델의 출력 결과를 가중 평균하여 최종 딥보이스 확률을 계산할 수 있다.

[96] 먼저, 시스템(400)은 음성 데이터에서 배경 소음, 호흡 소리, 입술 소리 등 미세 잡음을 분리한다. 이러한 잡음은 딥보이스 판별에 방해가 될 수 있으므로, 정확한 분석을 위해 제거하거나 감쇠시킨다.

[97] 시스템(400)은 음소 지속 시간, 공명 특징, 조음 방식 등 발음 패턴을 분석한다. 딥보이스는 사람의 발음과 미세하게 다를 수 있으므로, 이러한 발음

[98] 패턴의 차이를 감지하여 딥보이스 여부를 판별하는 데 활용한다. 시스템(400)은 음성 휴지 시간, 즉 단어 사이의 간격과 문장 사이의 간격을 분석한다. 딥보이스는 사람과 달리 호흡이 필요 없으므로, 휴지 시간 패턴이 사람과 다르게 나타날 수 있다.

[99] 시스템(400)은 발화 내용, 대화 주제, 상황 정보 등 문맥 정보를 분석한다. 딥보이스는 문맥에 맞지 않는 발화를 하거나, 대화 주제를 제대로 이해하지 못하는 경우가 있을 수 있다. 이러한 문맥 정보를 활용하여 딥보이스 여부를 판별한다. 추출된 특징들은 앙상블 AI 모델에 입력되어 딥보이스 여부를 판별한다.

[100] 앙상블 AI 모델은 딥러닝 기반의 음향 모델, 자연어 처리 기반의 언어 모델, 통계 기반의 패턴 인식 모델을 결합한 것이다.

[101] 음향 모델은 음성 데이터의 음향적 특징을 분석하여 딥보이스 여부를 판별한다. 언어 모델은 발화 내용의 문법적, 의미적 오류를 분석하여 딥보이스 여부를 판별한다. 패턴 인식 모델은 음성 데이터의 통계적 패턴을 분석하여 딥보이스 여부를 판별한다. 각 모델은 독립적으로 딥보이스 여부를 판별하고, 시스템(400)은 각 모델의 출력 결과를 가중 평균하여 최종 딥보이스 확률을 계산한다. 가중치는 각 모델의 성능에 따라 결정되며, 성능이 좋은 모델에 더 높은 가중치를 부여한다. 이를 통해 각 모델의 장점을 결합하고 단점을 보완하여, 딥보이스 판별의 정확도를 높인다.

[102] 동작 S630에서, 시스템(400)은 딥보이스로 판별된 경우 경고를 발생시키고 발신자의 신원을 확인할 수 있는 추가 질문을 제시한다. 구체적으로, 시스템(400)은 산출된 딥보이스 확률이 기 설정된 임계값(예: 사용자 설정값 또는 시스템 기본값) 이상인 경우 해당 음성을 딥보이스로 판별한다. 딥보이스로 판별되면, 시스템은 즉시 사용자 단말에 "주의! 딥보이스 의심 음성이 감지되었습니다."라는 시각적 경고 메시지를 표시함과 동시에, 진동 또는 경고음을 발생시켜 사용자가 위험을 즉각 인지할 수 있도록 한다. 이후, 시스템(400)은 발신자가 실제 지인인지 확인하기 위해 "발신자가 본인의 생년월일을 알고 있습니까? " 또는 "최근 발

신자와 나눈 대화 주제는 무엇입니까? "와 같은 추가 질문을 제시한다. 만약 발신자가 이에 대해 정확한 답변(예: "1990년 1월 1일입니다")을 하지 못하거나 머뭇거릴 경우, 시스템은 이를 딥보이스 판별의 추가적인 근거로 활용할 수 있다.

[103] 동작 S640에서, 시스템(400)은 수집된 사용자 응답을 분석하여 딥보이스 확률을 재평가할 수 있다. 시스템(400)은 상기 사용자의 음성 또는 텍스트 형식의 응답을 수집하고, 상기 수집된 사용자 응답을 분석하여 딥보이스 확률을 재평가하며, 상기 재평가된 딥보이스 확률을 "딥보이스 확률: XX%" 와 같이 수치 또는 그래프 형태로 사용자에게 제공할 수 있다.

[104] 시스템은 수집된 사용자 응답을 분석하여 딥보이스 확률을 재평가한다. 사용자가 음성 또는 텍스트 형식으로 응답하면, 시스템은 이 응답을 수집하고 분석한다. 예를 들어, 사용자가 추가 질문에 대해 "저는 생년월일을 알고 있습니다" 라고 대답하면, 시스템은 이 응답을 처리하여 발신자의 신원을 확인하는 데 필요한 정보를 얻는다.

[105] 이 과정에서 시스템은 사용자의 응답의 일관성을 평가한다. 만약 사용자가 "최근 대화 주제는 여행입니다"라고 대답했을 때, 시스템은 이전 대화 기록과 비교하여 이 응답이 신뢰할 수 있는지를 판단한다. 이후, 시스템은 수집된 사용자 응답을 바탕으로 딥보이스 확률을 재평가한다. 예를 들어, 초기 딥보이스 확률이 85%였다면, 사용자가 제공한 정확한 정보로 인해 이 확률이 50%로 낮아질 수 있다. 재평가된 딥보이스 확률은 "딥보이스 확률: XX%"와 같은 형식으로 수치 또는 그래프 형태로 사용자에게 제공된다. 이를 통해 사용자는 현재 상황을 명확히 이해할 수 있다.

[106] 사용자는 시스템(400)이 제시한 추가 질문에 음성 또는 텍스트 형식으로 응답할 수 있다. 시스템(400)은 사용자의 응답을 수집하고, 자연어 처리 기술을 이용하여 응답 내용을 분석한다. 분석 결과를 토대로 딥보이스 확률을 재평가한다. 예를 들어, 사용자가 질문에 정확하게 답변하지 못하거나, 답변 내용이 모호하거나 일관성이 없는 경우 딥보이스 확률을 높인다. 반대로, 사용자가 질문에 정확하고 명확하게 답변하는 경우 딥보이스 확률을 낮춘다. 재평가된 딥보이스 확률은 "딥보이스 확률: XX%" 와 같이 수치로 표시하거나, 막대 그래프 또는 원형 차트와 같은 그래프 형태로 시각적으로 표현하여 사용자에게 제공한다. 이를 통해 사용자는 딥보이스 여부를 보다 직관적으로 판단할 수 있다.

[107]

[108] 단순한 정답 일치 여부뿐만 아니라, 시스템은 수신된 사용자의 응답 음성 신호에서 '생체 활성 반응(Liveness Response)'을 분석하여 확률 재평가에 반영한다. 구체적으로, 질문 제시 시점부터 응답 시작 시점까지의 지연 시간(Latency)을 밀리초(ms) 단위로 측정하고, 응답 음성 파형의 지터(Jitter) 및 쉬머(Shimmer) 변동률을 산출한다. 만약 응답 지연 시간이 일반적인 인간의 반응 속도 범위(예: 200~500ms)를 벗어나거나, 파형의 지터 변동률이 인공 합성음 특유의 평탄한 패턴(예: 0.5% 미만)을 보일 경우, 답변 내용이 정답이라 하더라도 딥보이스 확률에 페널

티 가중치(예: +0.2)를 가산하여 최종 확률을 상향 업데이트하는 피드백 루프를 수행한다.

[109]

[110] 동작 S650에서, 시스템(400)은 주기적으로 딥보이스 검출 모델을 업데이트할 수 있다. 시스템(400)은 상기 사용자의 선택에 따라, 딥보이스 의심 통화를 즉시 차단하며, 상기 사용자의 선택에 따라, 통화 내용을 녹음하여 사용자의 클라우드 저장소에 저장하고, 녹음 파일을 분석하여 딥보이스 여부를 판단하는 데 활용하며, 주기적으로 딥보이스 검출 모델을 업데이트할 수 있다. 업데이트는 최신 딥보이스 생성 기술을 반영한 새로운 데이터를 이용하여 모델을 재훈련하는 것을 포함할 수 있다. 시스템(400)은 업데이트된 모델을 사용자 단말에 자동으로 배포하고, 발신자 번호, 통화 시간, 딥보이스 확률, 사용자 응답, 녹음 파일 중 적어도 하나를 포함하는 딥보이스 의심 음성 데이터와 관련된 정보를 암호화하여 경찰청 사이버 안전국에 자동으로 신고하도록 제어할 수 있다.

[111] 시스템은 주기적으로 딥보이스 검출 모델을 업데이트한다. 사용자가 특정 상황에서 딥보이스 의심 통화를 즉시 차단하기로 선택하면, 시스템은 해당 통화를 차단한다. 예를 들어, 사용자가 발신자의 음성이 의심스럽다고 판단하여 통화를 중단하고자 할 때, 시스템은 이를 즉시 실행한다.

[112] 또한, 사용자가 통화 내용을 녹음하여 클라우드 저장소에 저장하도록 설정할 경우, 시스템은 통화 내용을 녹음하고 이를 사용자의 클라우드에 저장한다. 이후, 저장된 녹음 파일은 딥보이스 여부를 판단하는 데 활용된다. 예를 들어, 시스템은 이 녹음 파일을 분석하여 딥보이스 확률을 다시 계산할 수 있다.

[113] 주기적으로 딥보이스 검출 모델을 업데이트하는 과정은 최신 딥보이스 생성 기술을 반영한 새로운 데이터를 이용하여 모델을 재훈련하는 것을 포함한다. 예를 들어, 시스템은 최신 음성 합성 기술에 대한 데이터를 수집하고 이를 기반으로 모델을 업데이트하여 성능을 향상시킨다. 이러한 업데이트 과정이 완료되면, 시스템은 업데이트된 모델을 사용자 단말에 자동으로 배포한다.

[114] 마지막으로, 시스템은 발신자 번호, 통화 시간, 딥보이스 확률, 사용자 응답, 녹음 파일 중 적어도 하나를 포함하는 딥보이스 의심 음성 데이터와 관련된 정보를 암호화하여 경찰청 사이버 안전국에 자동으로 신고한다. 예를 들어, 시스템은 "발신자 번호: 010-1234-5678, 통화 시간: 2024년 10월 17일 15시 30분, 딥보이스 확률: 85%"와 같은 정보를 포함하여 신고를 진행할 수 있다.

[115] 사용자는 딥보이스 의심 통화를 즉시 차단하거나, 통화 내용을 녹음하여 나중에 확인할 수 있다. 녹음된 통화 내용은 사용자의 클라우드 저장소에 안전하게 저장된다. 녹음 파일은 딥보이스 여부를 판단하는 데 활용될 수 있으며, 향후 딥보이스 검출 모델을 개선하는 데에도 사용될 수 있다. 시스템(400)은 주기적으로 딥보이스 검출 모델을 업데이트한다. 업데이트는 최신 딥보이스 생성 기술을 반영한 새로운 데이터를 이용하여 모델을 재훈련하는 것을 포함한다. 예를 들어, 새로운 딥보이스 생성 기술이 등장하면, 해당 기술로 생성된 딥보이스 음성 데

이터를 수집하여 모델을 재훈련한다. 또한, 사용자들이 녹음한 딥보이스 의심 음성 데이터를 분석하여 모델을 개선할 수 있다. 업데이트된 모델은 사용자 단말에 자동으로 배포된다. 시스템(400)은 발신자 번호, 통화 시간, 딥보이스 확률, 사용자 응답, 녹음 파일 중 적어도 하나를 포함하는 딥보이스 의심 음성 데이터와 관련된 정보를 수집한다. 수집된 정보는 암호화하여 개인 정보를 보호하고, 경찰청 사이버 안전국에 자동으로 신고한다. 이를 통해 딥보이스 범죄 예방 및 수사에 기여할 수 있다.

[116] 도 7은 일 실시예에 따른 생성형 인공지능에 의해 생성된 딥보이스를 검출하고 관련 사고를 예방하는 인공지능 솔루션 제공 방법의 흐름도를 나타낸 것이다.

[117] 도 7의 순서도에서 프로세스 단계들, 방법 단계들, 알고리즘들 등이 순차적인 순서로 설명되었지만, 그러한 프로세스들, 방법들 및 알고리즘들은 임의의 적합한 순서로 작동하도록 구성될 수 있다. 다시 말하면, 본 발명의 다양한 실시예들에서 설명되는 프로세스들, 방법들 및 알고리즘들의 단계들이 본 발명에서 기술된 순서로 수행될 필요는 없다. 또한, 일부 단계들이 비동시적으로 수행되는 것으로서 설명되더라도, 다른 실시예에서는 이러 한 일부 단계들이 동시에 수행될 수 있다. 또한, 도면에서의 묘사에 의한 프로세스의 예시는 예시된 프로세스 가 그에 대한 다른 변화들 및 수정들을 제외하는 것을 의미하지 않으며, 예시된 프로세스 또는 그의 단계들 중 임의의 것이 본 발명의 다양한 실시예들 중 하나 이상에 필수적임을 의미하지 않으며, 예시된 프로세스가 바람 직하다는 것을 의미하지 않는다.

[118] 동작 S710에서, 시스템(예: 도 4의 시스템(400))은 프로세서(예: 도 4의 프로세서(420))의 제어 하에, 아날로그 음성 신호를 수집할 수 있다. 시스템(400)은 분산 제어 영역에 공간적으로 분리되어 배치된 복수의 어레이 마이크를 통해 음원을 획득하여 아날로그 음성 신호를 수집할 수 있다. 각 어레이 마이크는 소리의 방향을 감지할 수 있도록 배열된 복수의 마이크 유닛을 포함할 수 있다.

[119] 프로세서(420)는 분산 제어 영역에 공간적으로 분리되어 배치된 복수의 어레이 마이크를 제어하여 아날로그 음성 신호를 수집한다. 즉, 여러 개의 마이크를 사용하여 소리를 녹음함으로써, 단일 마이크를 사용하는 것보다 더 풍부하고 정확한 음성 정보를 얻는다.

[120] 각 어레이 마이크는 소리의 방향을 감지할 수 있도록 배열된 복수의 마이크 유닛을 포함한다. 이러한 마이크 배열을 통해 소리가 어느 방향에서 오는지 파악하고, 해당 방향의 소리를 더욱 선명하게 수집할 수 있다. 예를 들어, 여러 개의 마이크 유닛에서 수집된 신호의 시간 차이를 분석하여 음원의 위치를 파악하는 빔포밍 기술을 사용할 수 있다. 이러한 공간적인 정보를 활용하여, 시스템(400)은 주변 소음을 효과적으로 제거하고, 원하는 음성 신호만 선택적으로 수집할 수 있다.

[121] 동작 S720에서, 시스템(400)은 음성 활동 검출 알고리즘을 이용하여 음성 구간을 분할할 수 있다. 시스템(400)은 아날로그-디지털 변환기를 이용하여 상기 수

집된 아날로그 음성 신호를 실시간으로 디지털 음성 신호로 변환하며, 샘플링 레이트 및 비트 깊이를 조절하고, 적응형 필터링, 스펙트럴 감산, 빔포밍 중 적어도 하나를 이용하여 상기 디지털 음성 신호에서 잡음을 제거하고, 자동 이득 제어를 이용하여 음량을 조절하며, 음성 활동 검출 알고리즘을 이용하여 음성 구간을 분할할 수 있다.

[122] 시스템은 음성 활동 검출 알고리즘을 이용하여 음성 구간을 분할한다. 이 단계에서 시스템은 아날로그-디지털 변환기를 이용하여 수집된 아날로그 음성 신호를 실시간으로 디지털 음성 신호로 변환한다. 예를 들어, 사용자가 "안녕하세요"라고 말할 때, 시스템은 이 음성을 디지털 신호로 변환하면서 샘플링 레이트와 비트 깊이를 조절하여 음질을 최적화한다. 샘플링 레이트가 높을수록 음성이 보다 세밀하게 재현되며, 비트 깊이가 깊을수록 음색의 다양성을 더할 수 있다.

[123] 이후, 시스템은 적응형 필터링, 스펙트럴 감산, 빔포밍 중 적어도 하나를 이용하여 디지털 음성 신호에서 잡음을 제거한다. 예를 들어, 배경 소음이 많은 환경에서 적응형 필터링을 적용하여 사용자 음성만을 선명하게 추출할 수 있다.

[124] 또한, 시스템은 자동 이득 제어를 이용하여 음량을 조절한다. 이는 사용자의 음성이 너무 작거나 클 경우 자동으로 조정하여 일관된 음량을 유지하는 기능이다. 마지막으로, 음성 활동 검출 알고리즘을 이용하여 음성 구간을 분할한다. 이 알고리즘은 음성의 시작과 끝을 감지하여, 음성이 없는 구간과 음성이 있는 구간을 명확히 구분한다. 예를 들어, 사용자가 "안녕하세요"라고 말한 후 잠시 멈추는 경우, 시스템은 이 구간을 음성 구간으로 인식하고, 이후의 구간을 음성이 없는 구간으로 식별한다.

[125] 시스템(400)은 아날로그-디지털 변환기(ADC)를 이용하여 수집된 아날로그 음성 신호를 실시간으로 디지털 음성 신호로 변환한다. 이때, 샘플링 레이트와 비트 깊이를 조절하여 디지털 음성 신호의 품질을 결정한다. 샘플링 레이트는 1초에 몇 번 아날로그 신호를 측정하는지를 나타내며, 비트 깊이는 각 샘플의 진폭을 표현하는 데 사용되는 비트 수를 나타낸다. 디지털 음성 신호에서 잡음을 제거하기 위해 다양한 기술을 적용한다. 적응형 필터링은 주변 환경의 잡음 특성에 따라 필터의 계수를 자동으로 조정하여 잡음을 제거하는 기술이다. 스펙트럴 감산은 잡음 스펙트럼을 추정하고, 이를 원래 신호의 스펙트럼에서 빼서 잡음을 제거하는 기술이다. 빔포밍은 어레이 마이크를 사용하여 특정 방향의 소리를 강화하고 다른 방향의 소리를 감쇠시키는 기술이다.

[126] 시스템(400)은 자동 이득 제어(AGC)를 이용하여 음량을 조절한다.

[127] AGC는 입력 신호의 크기에 따라 자동으로 증폭 또는 감쇠하여 출력 신호의 크기를 일정하게 유지하는 기술이다.

[128] 시스템(400)은 음성 활동 검출(VAD) 알고리즘을 이용하여 음성 구간을 분할한다. VAD 알고리즘은 음성 신호의 에너지, 주파수 특성, 시간적 변화 등을 분석하여 음성이 존재하는 구간과 묵음 구간을 구분한다. 이를 통해 음성 인식 및 분석의 효율성을 높일 수 있다.

[129] 동작 S730에서, 시스템(400)은 각 모델의 출력 결과를 가중 평균하여 제1 딥보이스 확률을 계산할 수 있다. 시스템(400)은 전처리된 디지털 음성 신호에 서 프레임 단위로 멜 주파수 켑스트럼 계수(MFCC), 선형 예측 코딩(LPC) 계수, 피치, 포먼트, 스펙트로그램을 포함하는 음향 특징을 추출하고, 프레임 크기 및 프레임 이동 간격을 조절하고, 딥러닝 기반의 음향 모델, 자연어 처리 기반의 언어 모델, 및 통계 기반의 패턴 인식 모델을 포함하는 앙상블 AI 모델인 제1 생성형 AI 모델에 상기 추출된 음향 특징을 입력하여 각 모델의 출력 결과를 가중 평균하여 제1 딥보이스 확률을 계산할 수 있다.

[130] 시스템(400)은 각 모델의 출력 결과를 가중 평균하여 제1 딥보이스 확률을 계산한다. 이 과정에서 시스템은 전처리된 디지털 음성 신호에서 여러 음향 특징을 추출한다. 예를 들어, 시스템은 프레임 단위로 멜 주파수 켑스트럼 계수(MFCC), 선형 예측 코딩(LPC) 계수, 피치, 포먼트, 스펙트로그램을 포함하는 음향 특징을 추출한다. 각 특징은 음성을 다양한 방식으로 분석하는 데 사용된다. 예를 들어, MFCC는 음성의 주파수 성분을 나타내며, LPC는 음색의 특성을 모델링한다.

[131] 피치는 음성의 높낮이를 나타내고, 포먼트는 특정 모음의 주파수 대역을 나타낸다.

[132] 스펙트로그램은 시간에 따른 주파수 변화를 시각적으로 나타내어, 발음의 패턴을 분석하는 데 유용하다.

[133] 시스템(400)은 프레임 크기 및 프레임 이동 간격을 조절하여 각 음향 특징을 최적화한다. 예를 들어, 프레임 크기를 20ms로 설정하고, 이동 간격을 10ms로 설정하여, 짧은 시간 간격으로 세밀한 분석을 수행할 수 있다.

[134] 이후, 시스템(400)은 딥러닝 기반의 음향 모델, 자연어 처리 기반의 언어 모델, 및 통계 기반의 패턴 인식 모델을 포함하는 앙상블 AI 모델인 제1 생성형 AI 모델에 추출된 음향 특징을 입력한다. 각 모델은 입력된 음향 특징을 기반으로 딥보이스 여부를 판단하며, 이때 각 모델의 출력 결과를 가중 평균하여 제1 딥보이스 확률을 계산한다. 예를 들어, 음향 모델이 75%, 언어 모델이 80%, 패턴 인식 모델이 70%의 확률로 딥보이스라고 판단할 경우, 시스템은 이를 종합하여 최종 확률을 계산하게 된다.

[135] 시스템(400)은 전처리된 디지털 음성 신호를 프레임 단위로 나누고, 각 프레임에서 멜 주파수 켑스트럼 계수(MFCC), 선형 예측 코딩(LPC) 계수, 피치, 포먼트, 스펙트로그램을 포함하는 음향 특징을 추출한다. 프레임 크기와 프레임 이동 간격을 조절하여 분석의 정밀도를 조정할 수 있다. 추출된 음향 특징은 앙상블 AI 모델인 제1 생성형 AI 모델에 입력된다. 제1 생성형 AI 모델은 딥러닝 기반의 음향 모델, 자연어 처리 기반의 언어 모델, 통계 기반의 패턴 인식 모델을 포함한다. 각 모델은 입력된 음향 특징을 분석하여 딥보이스 여부를 판별한다.

[136] 시스템(400)은 각 모델의 출력 결과를 가중 평균하여 제1 딥보이스 확률을 계산한다. 가중치는 각 모델의 성능에 따라 결정되며, 성능이 좋은 모델에 더 높은 가

중치를 부여한다. 이를 통해 각 모델의 장점을 결합하고 단점을 보완하여 딥보이스 판별의 정확도를 높인다.

[137] 동작 S740에서, 시스템(400)은 시계열 분석 모델인 제2 생성형 AI 모델에 입력하여 제2 딥보이스 확률을 계산할 수 있다. 시스템(400)은 추출된 음향 특징을 이용하여 음성 신호의 피치 변화, 에너지 변화, 발음 속도 변화, 포먼트 변화, 스펙트럼 변화 중 적어도 하나를 포함하는 시간적 변화를 분석하고, 이를 순환 신경망(RNN) 기반 모델, 장단기 메모리(LSTM) 기반 모델, 및 게이트 순환 유닛(GRU) 기반 모델 중 적어도 하나를 포함하는 시계열 분석 모델인 제2 생성형 AI 모델에 입력하여 제2 딥보이스 확률을 계산할 수 있다.

[138] 시스템은 최종 딥보이스 확률을 계산한다. 이 과정에서 시스템은 가중 평균, 규칙 기반 결합, 머신러닝 기반 결합 중 적어도 하나를 이용하여 제1 딥보이스 확률과 제2 딥보이스 확률을 융합한다. 예를 들어, 제1 딥보이스 확률이 70%이고 제2 딥보이스 확률이 80%라면, 시스템은 이 두 확률을 가중 평균하여 최종 확률을 75%로 계산할 수 있다. 이 때, 최종 딥보이스 확률은 사용자 설정 가능하고 상황에 따라 동적으로 조절될 수 있는 기 설정된 임계값 이상인 경우, 시스템은 해당 음성을 딥보이스로 판별하고 경보 모드를 시작한다. 예를 들어, 사용자가 설정한 임계값이 75%일 경우, 최종 확률이 이 기준을 초과하면 시스템은 즉시 경고를 발령하게 된다.

[139] 시스템(400)은 추출된 음향 특징을 이용하여 음성 신호의 피치 변화, 에너지 변화, 발음 속도 변화, 포먼트 변화, 스펙트럼 변화 등 시간에 따른 변화 패턴을 분석한다. 딥보이스는 사람의 음성과 달리 시간적 변화 패턴에서 미세한 차이를 보일 수 있다. 예를 들어, 딥보이스는 사람보다 피치 변화가 단조롭거나, 발음 속도가 일정하게 유지되는 경향을 보일 수 있다.

[140] 분석된 시간적 변화 정보는 시계열 분석 모델인 제2 생성형 AI 모델에 입력된다. 제2 생성형 AI 모델은 순환 신경망(RNN) 기반 모델, 장단기 메모리(LSTM) 기반 모델, 게이트 순환 유닛(GRU) 기반 모델 중 적어도 하나를 포함한다.

[141] RNN, LSTM, GRU는 모두 시계열 데이터 분석에 특화된 딥러닝 모델로, 이전 시간의 정보를 기억하여 현재 시간의 데이터를 분석하는 데 활용한다.

[142] 제2 생성형 AI 모델은 입력된 시간적 변화 정보를 분석하여 제2 딥보이스 확률을 계산한다. 즉, 음성 신호의 시간적 변화 패턴이 딥보이스에서 나타나는 패턴과 얼마나 유사한지 분석하여 딥보이스 여부를 판별한다.

[143] 동작 S750에서, 시스템(400)은 최종 딥보이스 확률을 계산할 수 있다. 시스템(400)은 가중 평균, 규칙 기반 결합, 머신러닝 기반 결합 중 적어도 하나를 이용하여 상기 상기 제1 딥보이스 확률과 상기 제2 딥보이스 확률을 융합하여 최종 딥보이스 확률을 계산하고, 상기 최종 딥보이스 확률이 사용자 설정 가능하고 상황에 따라 동적으로 조절될 수 있는 기 설정된 임계값 이상인 경우, 딥보이스로 판별하고 경보 모드를 시작할 수 있다.

[144]

[145] 도 8은 본 발명의 일 실시예에 따른 신호 품질(SNR) 기반의 적응형 가중치 할당 및 리소스 최적화를 이용한 딥보이스 검출 방법을 나타내는 순서도이다.

[146] 단계 S810에서, 시스템(400)은 입력된 음성 데이터로부터 딥보이스 판별에 필요한 복합적인 특징 벡터를 추출한다. 이때 추출되는 특징은, 주파수 도메인의 특성을 나타내는 멜 스펙트로그램(Mel-Spectrogram), 시간적 변화 및 성도(Vocal Tract)의 특성을 나타내는 MFCC 및 LPC 계수, 그리고 발화의 의미적 맥락을 포함하는 텍스트 데이터를 포함한다.

[147] 단계 S820에서, 시스템(400)은 추출된 특징 벡터를 복수의 AI 모델을 포함하는 앙상블 모델에 입력하여 개별 확률을 산출한다. 본 발명의 일 실시예에 따르면, 상기 앙상블 모델은, 멜 스펙트로그램을 분석하는 CNN(Convolutional Neural Network) 기반의 제1 AI 모델(음향), 텍스트의 문맥을 분석하는 트랜스포머(Transformer) 기반의 제2 AI 모델(언어), 및 피치/에너지 통계치를 분석하는 제3 AI 모델(패턴)로 구성될 수 있다. 특히, 이 과정에서 시스템(400)은 '동적 리소스 스위칭(Dynamic Resource Switching)'을 수행하여 하드웨어 자원을 최적화한다. 구체적으로, 프로세서는 평상시에는 연산 부하가 적은 제3 AI 모델(패턴 인식)만을 활성화하여 1차 스크리닝을 수행한다. 1차 분석 결과 딥보이스 확률이 예비 임계값을 초과하는 의심 구간이 탐지되지 않는다면, 대용량 메모리를 점유하는 제1 AI 모델(CNN) 및 제2 AI 모델(Transformer)은 비활성 상태(Unloaded)를 유지하여 전력 소모와 메모리 점유율을 최소화한다. 만약 의심 구간이 탐지되면, 프로세서는 즉시 상기 제1 및 제2 AI 모델을 활성 메모리(Active Memory)에 로드(Load)하여 정밀 분석을 수행한다.

[148] 단계 S830에서, 시스템(400)은 입력된 음성 데이터의 신호 품질을 분석하여 신호 대 잡음비(SNR)를 측정한다. 이는 주변 소음이 심한 환경이나 통신 상태가 좋지 않은 환경에서 특정 모델의 오작동을 방지하기 위함이다.

[149] 단계 S840에서, 시스템(400)은 측정된 SNR 값에 기초하여 각 AI 모델에 적용할 가중치를 실시간으로 변동시키는 '적응형 가중치 할당(Adaptive Weight Allocation)'을 수행한다. 예를 들어, 측정된 SNR 값이 기설정된 기준치(예: 10dB) 미만인 저품질 신호(Noisy)인 경우, 미세한 주파수 패턴을 분석하는 제1 AI 모델(음향)의 신뢰도가 하락할 수 있다. 이에 따라 시스템은 제1 AI 모델의 가중치를 감소시키고(예: $w_1 = 0.2$), 상대적으로 소음의 영향을 덜 받는 제2 AI 모델(언어/문맥)의 가중치를 증가시키는(예: $w_2 = 0.5$) 보정을 수행한다.

[150] 단계 S850에서, 시스템(400)은 결정된 동적 가중치와 각 모델의 개별 확률을 연산하여 최종 딥보이스 확률(P_{final})을 산출한다. 이때 적용되는 수식은, 각 모델의 확률(p_i)과 SNR에 따라 변동되는 가중치(w_i)의 가중 평균 연산

[151]
$$P_{final} = \frac{\sum(w_i \times p_i)}{\sum w_i}$$

[152] 을 따른다. 이를 통해 환경 변수에 강인한 정밀 탐지가 가능하다.

[153] 마지막으로, 최종 딥보이스 확률이 임계값을 초과하는 경우(S860의 예), 시스템은 즉시 경보를 발생시키거나 신고 절차를 수행한다.

[154] 나아가, 본 발명의 또 다른 실시예로서 '생체 활성 감지(Liveness Detection) 피드백' 과정이 단계 S860 이후 또는 병행하여 수행될 수 있다. 시스템(400)은 딥보이스가 의심될 때 사용자 단말을 통해 발신자에게 추가 질문(Challenge)을 제시하고, 이에 대한 발신자의 응답(Response)을 분석한다. 이때 시스템은 답변의 정답 여부뿐만 아니라, 질문 시점부터 응답 시점까지의 지연 시간(Latency) 및 응답 음성 파형의 지터(Jitter)와 쉬머(Shimmer) 변동률을 밀리초(ms) 단위로 측정한다.만약 측정된 지연 시간이 인간의 평균 반응 속도(예: 200~500ms)를 벗어나거나(AI 연산 지연), 지터 변동률이 0.5% 미만으로 지나치게 평탄한(인공 합성음 특유의) 패턴을 보일 경우, 시스템은 이를 '기계적 발화'로 간주한다. 이에 따라 시스템은 답변 내용이 정확하더라도 최종 딥보이스 확률에 페널티 가중치(예: +0.2)를 가산하여 확률을 상향 업데이트(Update)함으로써, 교묘하게 학습된 딥보이스 공격을 효과적으로 차단한다.

[155]

[156] 가중 평균, 규칙 기반 결합, 머신러닝 기반 결합 중 적어도 하나를 이용하여 제1 딥보이스 확률과 제2 딥보이스 확률을 융합하여 최종 딥보이스 확률을 계산한다. 가중 평균은 각 확률에 가중치를 부여하여 평균을 계산하는 방식이다.

[157] 규칙 기반 결합은 미리 정의된 규칙에 따라 두 확률을 결합하는 방식이다. 예를 들어, "두 확률 중 하나라도 0.8 이상이면 최종 확률을 0.9로 설정한다" 와 같은 규칙을 사용할 수 있다. 머신러닝 기반 결합은 머신러닝 모델을 사용하여 두 확률을 결합하는 방식이다. 예를 들어, 두 확률을 입력으로 받아 최종 확률을 출력하는 분류 모델을 학습시킬 수 있다. 최종 딥보이스 확률이 사용자 설정 가능하고 상황에 따라 동적으로 조절될 수 있는 기 설정된 임계값 이상인 경우, 딥보이스로 판별한다. 임계값은 사용자가 직접 설정하거나, 시스템이 상황에 따라 자동으로 조절할 수 있다. 예를 들어, 사용자가 보안 수준을 높게 설정하면 임계값이 낮아져 딥보이스로 판별될 확률이 높아진다. 반대로, 사용자가 편의성을 중시하면 임계값이 높아져 딥보이스로 판별될 확률이 낮아진다.

[158] 일 실시예에 따르면, 시스템(400)은 경보 모드 시작 시, "딥보이스가 감지되었습니다.", "주의하십시오.", "경고! 딥보이스 사용 중단"의 메시지를 포함할 수 있고 사용자 설정 가능한 미리 설정된 경고 음성을 재생하고 디지털 오디오 전력 증폭을 통해 라우드 스피커로 출력하고, 경고 음성 출력과 동시에, 텍스트, 이미지, 애니메이션 등 다양한 형태로 표시될 수 있는 시각적 경고 메시지를 디스플레이에 표시하고, 사용자의 주의를 끌 수 있는 다양한 패턴으로 진동 또는 경고음을 발생시키며, 네트워크를 통해 관리 센터 호스트에 딥보이스 의심 음성 데이터, 발신자 정보, 딥보이스 확률, 현재 시간, 위치 정보, 마이크 ID, 및 발생 시간 정보를 포함하는 경보 시스템을 전송할 수 있다.

[159] 경보 모드 시작 시, "딥보이스가 감지되었습니다.", "주의하십시오.", "경고! 딥보이스 사용 중단" 과 같은 메시지를 포함하는 경고 음성을 재생한다. 경고 음성은 사용자 설정 가능한 미리 설정된 음성이 사용될 수 있다. 디지털 오디오 전력 증폭을 통해 경고 음성을 라우드 스피커로 출력하여 사용자 주변에 있는 사람들에게도 경고 메시지를 전달한다. 경고 음성 출력과 동시에, 텍스트, 이미지, 애니메이션 등 다양한 형태로 표시될 수 있는 시각적 경고 메시지를 디스플레이에 표시한다. 예를 들어, 딥보이스 경고 이미지와 함께 "딥보이스가 감지되었습니다!" 라는 텍스트 메시지를 화면에 표시할 수 있다. 시스템(400)은 사용자의 주의를 끌기 위해 다양한 패턴으로 진동 또는 경고음을 발생시킨다. 예를 들어, 짧게 여러 번 진동하거나, 높낮이가 변하는 경고음을 발생시킬 수 있다.

[160] 시스템(400)은 네트워크를 통해 관리 센터 호스트에 딥보이스로 의심되는 음성 데이터, 발신자 정보, 딥보이스 확률, 현재 시간, 위치 정보, 마이크 ID, 및 발생 시간 정보를 포함하는 경보 신호(또는 경보 메시지)를 전송한다. 이를 통해 관리자는 딥보이스 발생 상황을 실시간으로 파악하고, 필요한 조치를 취할 수 있다.

[161] 일 실시예에 따르면, 시스템(400)은 상기 관리 센터 호스트에서 경보 정보를 수신하도록 제어하고, 경보 발생 위치와 함께 발신자의 성별, 연령, 출신 지역, 사용 언어, 음성 특징 등을 포함할 수 있는 발신자의 음성 프로필 정보를 경보 표시 모듈에 표시하고, 상기 경보 정보 수신 후, 현장 모니터링을 시작하여 CCTV 영상, 온도, 습도, 조도, 움직임 감지 센서 등에서 수집된 정보를 포함할 수 있는 센서 데이터, 및 사용자의 위치, 권한, 접근 기록 등을 포함할 수 있는 사용자 정보를 종합적으로 분석하여 현장 정보를 확인하고, 상기 현장 정보 확인 후, 마이크 및 분산 제어 영역을 통해 양방향 통신을 지원하며 잡음 제거 및 음성 향상 기술이 적용된 음성 인터콤 재확인을 수행할 수 있다.

[162] 시스템(400)은 관리 센터 호스트에서 전송된 경보 정보를 수신한다.

[163] 경보 정보에는 딥보이스 발생 위치, 발신자의 음성 프로필 정보 등이 포함된다. 발신자의 음성 프로필 정보는 발신자의 성별, 연령, 출신 지역, 사용 언어, 음성 특징 등을 포함할 수 있다. 이러한 정보는 딥보이스 사용자를 식별하고 추적하는 데 도움을 줄 수 있다.

[164] 시스템(400)은 수신된 경보 정보를 경보 표시 모듈에 표시한다. 경보 표시 모듈은 딥보이스 발생 위치를 시각적으로 표시하고, 발신자의 음성 프로필 정보를 사용자에게 제공한다. 경보 정보 수신 후, 시스템(400)은 현장 모니터링을 시작한다. 현장 모니터링은 CCTV 영상, 온도, 습도, 조도, 움직임 감지 센서 등에 서 수집된 정보를 포함하는 센서 데이터, 및 사용자의 위치, 권한, 접근 기록 등을 포함하는 사용자 정보를 종합적으로 분석하여 현장 정보를 확인하는 과정이다.

[165] 시스템(400)은 수집된 현장 정보를 분석하여 딥보이스 발생 상황을 더욱 정확하게 파악한다. 예를 들어, CCTV 영상 분석을 통해 딥보이스가 사용된 위치와 시간을 확인하고, 움직임 감지 센서 데이터를 통해 딥보이스 사용자의 이동 경로

를 추적할 수 있다. 현장 정보 확인 후, 시스템(400)은 마이크 및 분산 제어 영역을 통해 양방향 통신을 지원하며 잡음 제거 및 음성 향상 기술이 적용된 음성 인터콤을 통해 재확인을 수행한다. 즉, 딥보이스가 발생한 현장에 있는 사람과 직접 대화하여 상황을 파악하고, 딥보이스 사용 여부를 최종적으로 확인한다.

[166] 일 실시예에 따르면, 시스템은 관리 센터 호스트에서 경보 정보를 수신하도록 제어한다. 이 과정에서 시스템은 발신자의 성별, 연령, 출신 지역, 사용 언어, 음성 특징 등을 포함할 수 있는 발신자의 음성 프로필 정보를 경보 표시 모듈에 표시한다. 예를 들어, 사용자가 통화 중에 의심스러운 음성을 감지했을 때, 시스템은 "발신자는 남성으로 추정되며, 30세로 보이고, 서울 출신이며, 한국어를 사용하고 있습니다"라는 정보를 화면에 표시하여 신속한 판단을 돕는다.

[167] 상기 경보 정보 수신 후, 시스템은 현장 모니터링을 시작한다. 이때 시스템은 CCTV 영상, 온도, 습도, 조도, 움직임 감지 센서 등에서 수집된 정보를 포함할 수 있는 센서 데이터를 이용하여 상황을 종합적으로 분석한다. 예를 들어, CCTV를 통해 특정 공간에서의 움직임을 감지하고, 온도와 습도 데이터를 통해 환경 조건을 파악하여 이상 징후를 탐지한다. 또한, 시스템은 사용자의 위치, 권한, 접근 기록 등을 포함할 수 있는 사용자 정보를 종합적으로 분석하여 현장 정보를 확인한다. 예를 들어, 특정 사용자가 권한 없이 접근한 이력이 있을 경우, 시스템은 이를 기반으로 경고를 발령할 수 있다.

[168] 현장 정보 확인 후, 시스템은 마이크 및 분산 제어 영역을 통해 양방향 통신을 지원하며 잡음 제거 및 음성 향상 기술이 적용된 음성 인터콤 재확인을 수행한다. 예를 들어, 사용자가 통화 중에 발생할 수 있는 배경 소음을 제거하여 상대방의 목소리를 더욱 선명하게 전달할 수 있다. 이를 통해 사용자와의 의사 소통이 원활하게 이루어지도록 한다.

[169] 일 실시예에 따르면, 시스템(400)은 상기 음성 인터콤 재확인 과정에서 실시간 음성 인터콤을 일시 중지하고 딥페이크 음성 사용자에게 음성 합성 기술을 이용하여 생성될 수 있으며 다양한 언어 및 음성으로 제공될 수 있는 "딥보이스 사용이 감지되었습니다. 법적 책임이 따를 수 있습니다."라는 경고 메시지를 전달하고, 상기 관리 센터에서 딥페이크 음성 사용자의 정보를 수신하고 딥페이크 음성 사용자의 위치 정보, 위험도, 대응 매뉴얼, 접근 경로, 주변 상황 정보, 및 관련 법규 정보를 포함하는 경보를 발령하여 보안 요원에게 알리며, 딥보이스 검출 결과를 기반으로 음성 사용 빈도, 시간, 장소, 대화 상대, 사용하는 단어, 감정 변화 등을 포함하는 사용자 행동 패턴을 분석하여 잠재적 위험 상황을 사전에 예측하고, 딥보이스 방지 교육, 보안 시스템 강화, 사용자 인증 절차 강화 등의 경고 및 예방 조치를 사용자 단말에게 제안하도록 제어할 수 있다.

[170] 음성 인터콤 재확인 과정에서 시스템(400)은 실시간 음성 인터콤을 일시 중지하고 딥페이크 음성 사용자에게 음성 합성 기술을 이용하여 생성된 경고 메시지를 전달한다. 경고 메시지는 "딥보이스 사용이 감지되었습니다. 법적 책임이 따

를 수 있습니다." 와 같은 내용을 포함하며, 다양한 언어 및 음성으로 제공될 수 있다.

[171] 시스템(400)은 관리 센터에서 딥페이크 음성 사용자의 정보를 수신한다. 딥페이크 음성 사용자의 정보는 위치 정보, 위험도, 대응 매뉴얼, 접근 경로, 주변 상황 정보, 및 관련 법규 정보를 포함한다. 시스템(400)은 수신된 정보를 바탕으로 경보를 발령하여 보안 요원에게 알린다. 시스템(400)은 딥보이스 검출 결과를 기반으로 음성 사용 빈도, 시간, 장소, 대화 상대, 사용하는 단어, 감정 변화 등을 포함하는 사용자 행동 패턴을 분석한다. 분석된 행동 패턴을 통해 잠재적 위험 상황을 사전에 예측한다. 예를 들어, 특정 시간대에 특정 장소에서 특정 단어를 사용하는 빈도가 높으면, 이를 잠재적 위험 상황으로 판단할 수 있다.

[172] 시스템(400)은 딥보이스 방지 교육, 보안 시스템 강화, 사용자 인증 절차 강화 등의 경고 및 예방 조치를 사용자 단말에게 제안한다. 예를 들어, 딥보이스 방지 교육 자료를 제공하거나, 비밀번호 변경, 2단계 인증 설정 등 보안 강화 조치를 권고할 수 있다.

[173] 일 실시예에 따르면, 시스템(400)은 센서를 통해 실시간으로 음성 데이터를 수집하고, 수집된 음성 데이터에 대해 잡음 제거, 음량 조절, 음성 구간 분할을 포함하는 전처리 과정을 수행하고, 종점 감지를 통해 상기 전처리된 음성 데이터에서 음성 세그먼트 및 잡음 세그먼트를 판별하고, 잡음 세그먼트를 제거하고, 상기 음성 세그먼트에 대해 해밍 윈도우를 사용하여 프레임화를 수행할 수 있다.

[174] 일 실시예에 따르면, 시스템(400)은 각 프레임에 대해 1차 고역 통과 필터를 사용하는 고주파수 프리엠퍼시스 처리를 수행하고, 상기 음성 세그먼트의 음성 데이터에 대해, 기본 음성 신호 주파수 스펙트럼에서 성대 여기로 인한 영향을 제거하고, 리프팅 방법을 통해 포먼트의 영향을 제거하여 성문 파형 신호를 획득하고, 선형 예측 코딩 및 이산 완전 또는 수집된 작품의 모델을 정확하게 사용하여 음향 모델을 구축하고, 최종적으로 리프팅을 사용하여 성문 파형 신호를 획득할 수 있다.

[175] 일 실시예에 따르면, 시스템(400)은 상기 성문 파형 신호에서 정규화된 진폭, 성문 폐쇄 시간, 피치 변화, 포먼트 변화, 스펙트로그램, 멜 주파수 켑스트럼 계수(MFCC)를 포함하는 특징 파라미터를 추출하고, 생성형 AI 모델에 의해 생성된 딥보이스 데이터를 포함하는 학습 데이터를 이용하여 훈련된 SVM 모델에 상기 추출된 특징 파라미터를 입력으로 제공하여 분류를 수행하고, 분류 결과를 이용하여 정상 음성, 딥보이스 의심, 딥보이스 확정 중 하나를 포함하는 화자의 상태를 판별하고, 화자 상태 라벨을 출력하며, 상기 화자 상태 라벨에 기반하여 딥보이스 의심 또는 딥보이스 확정 라벨 수신 시 경고 메시지 표시, 통화 차단, 녹음, 경찰청 사이버 안전국 신고 중 적어도 하나의 동작을 수행할 수 있다.

[176] 일 실시예에 따르면, 시스템은 음성 인터콤 재확인 과정에서 실시간 음성 인터콤을 일시 중지한다. 이때 시스템은 딥페이크 음성 사용자에게 음성 합성 기술을 이용하여 생성될 수 있는 경고 메시지를 전달한다. 예를 들어, 시스템은 "딥보

이스 사용이 감지되었습니다. 법적 책임이 따를 수 있습니다."라는 경고 메시지를 다양한 언어 및 음성으로 제공한다. 이는 사용자가 이해할 수 있도록 다국적 환경에서도 유용하게 쓰일 수 있다. 상기 경고 메시지가 전달된 후, 시스템은 관리 센터에서 딥페이크 음성 사용자의 정보를 수신한다. 이 정보에는 딥페이크 음성 사용자의 위치 정보, 위험도, 대응 매뉴얼, 접근 경로, 주변 상황 정보, 및 관련 법규 정보가 포함된다. 예를 들어, 사용자가 "위험도: 높음, 접근 경로: 3번 출입구"와 같은 정보를 통해 보안 요원에게 신속하게 대응할 수 있도록 한다.

[177] 이후 시스템은 보안 요원에게 경고를 발령하여 신속하게 상황에 대응할 수 있도록 한다. 시스템은 딥보이스 검출 결과를 기반으로 음성 사용 빈도, 시간, 장소, 대화 상대, 사용하는 단어, 감정 변화 등을 포함하는 사용자 행동 패턴을 분석하여 잠재적 위험 상황을 사전에 예측한다. 예를 들어, 사용자가 특정 시간에 자주 의심스러운 발언을 하는 패턴이 발견되면, 시스템은 이를 기록하여 위험 요소로 간주한다. 마지막으로, 시스템은 딥보이스 방지 교육, 보안 시스템 강화, 사용자 인증 절차 강화 등의 경고 및 예방 조치를 사용자 단말에게 제안하도록 제어한다. 예를 들어, 시스템은 "딥보이스 사용 방지를 위한 교육 프로그램에 참여하십시오"라는 알림을 사용자에게 보내어 예방 조치를 취할 수 있도록 유도한다. 이를 통해 사용자의 인식을 높이고, 잠재적인 위험을 줄이는 데 기여할 수 있다.

[178] 일 실시예에 따르면, 시스템(400)은 디바이스와 연관된 프로필 처리 요청에 응답하여 음향 신호를 방출하고, 사용자와 연관된 하나 이상의 깊이 부분의 고유한 윤곽으로부터 반사된 음향 신호를 기반으로 맞춤화된 일련의 에코 음향 신호를 수신하며, 분류 모델을 훈련시키기 위해 에코 음향 신호와 연관된 하나 이상의 영역 세그먼트를 추출하고, 추출된 하나 이상의 영역 세그먼트를 기반으로 분류 모델을 생성할 수 있다.

[179] 디바이스와 연관된 프로필 처리 요청에 응답하여, 시스템(400)은 스피커를 통해 특정 주파수를 가진 음향 신호를 방출한다. 이 음향 신호는 사용자의 신체, 특히 얼굴과 같은 하나 이상의 깊이 부분의 고유한 윤곽에서 반사되어 돌아온다. 시스템(400)은 마이크를 통해 반사된 음향 신호를 수신하고, 이를 맞춤화된 일련의 에코 음향 신호로 변환한다. 분류 모델은 사용자를 인증하기 위해 사용된다. 분류 모델을 훈련시키기 위해, 시스템(400)은 에코 음향 신호에서 하나 이상의 영역 세그먼트를 추출한다. 이러한 영역 세그먼트는 사용자의 고유한 신체적 특징을 나타내는 정보를 포함한다. 추출된 영역 세그먼트를 사용하여 분류 모델을 훈련시킨다.

[180] 예를 들어, 사용자의 코, 입, 귀 등의 깊이 정보가 담긴 에코 음향 신호에서 특징적인 패턴을 추출하고, 이를 기반으로 "사용자 A", "사용자 B" 등을 구분하는 분류 모델을 생성할 수 있다.

[181] 일 실시예에 따르면, 시스템(400)은 분류 모델을 기반으로 결합 특징 표현을 추출하며, 상기 결합 특징 표현은 이산 에포크 동안 동시에 얻어진 사용자의 음향 특징 및 시각적 랜드마크 특징을 포함하고, 결합 특징 표현의 예측에 사용되는

벡터 기반 분류 모델을 생성하고, 결합 특징 표현의 예측을 기반으로 결합 특징 표현이 에코-시그니처와 연관되어 있는지 여부를 결정하며, 사용자 프로필과 연관된 결합 특징 표현에는 이산 에포크에 대한 하나 이상의 깊이 부분의 고유한 윤곽과 관련된 추출된 랜드마크 좌표 및 음향 특징이 포함되고, 사용자 프로필과 연관된 결합 특징 표현을 생성할 때, 변경된 이산 에포크에 대한 사용자와 연관된 하나 이상의 깊이 부분의 고유한 윤곽과 관련된 합성된 음향 특징 및 증강된 랜드 마크 좌표로 결합 특징 표현을 보강할 수 있다.

[182] 시스템(400)은 훈련된 분류 모델을 기반으로 결합 특징 표현을 추출한다. 결합 특징 표현은 이산 에포크(discrete epoch), 즉 특정 시간 동안 동시에 얻어진 사용자의 음향 특징과 시각적 랜드마크 특징을 포함한다. 음향 특징은 앞서 설명한 에코 음향 신호에서 추출된 특징이다. 시각적 랜드마크 특징은 카메라를 통해 얻은 사용자 얼굴 이미지에서 눈, 코, 입 등의 위치 정보를 추출한 것이다.

[183] 시스템(400)은 결합 특징 표현을 예측하는 데 사용되는 벡터 기반 분류 모델을 생성한다. 이 분류 모델은 결합 특징 표현이 에코-시그니처(echo signature), 즉 사용자의 고유한 음향 및 시각적 특징과 일치하는지 여부를 판단한다. 사용자 프로필에는 이산 에포크에 대한 하나 이상의 깊이 부분의 고유한 윤곽과 관련된 추출된 랜드마크 좌표 및 음향 특징이 포함된다. 사용자 프로필을 생성할 때, 시스템(400)은 변경된 이산 에포크에 대한 사용자의 합성된 음향 특징 및 증강된 랜드마크 좌표를 사용하여 결합 특징 표현을 보강할 수 있다. 이는 시간이 지남에 따라 사용자의 외모나 음향 특징이 변할 수 있음을 고려하여 인증 정확도를 유지하기 위한 것이다.

[184] 일 실시예에 따르면, 시스템(400)은 변경된 이산 에포크에 대한 결합 특징 표현을 생성하기 위해 합성된 음향 특징 및/또는 증강된 랜드마크 좌표와 연관된 하나 이상의 변경된 벡터 특징을 추출할 수 있다. 컴퓨팅 디바이스와 연관된 프로필 처리 요청은 원본 사용자 프로필의 초기 등록 또는 컴퓨팅 디바이스 접근을 위한 현재 에포크에 대한 사용자 프로필 인증을 포함할 수 있다. 컴퓨팅 디바이스와 연관된 프로필 처리 요청은 또한 현재 에포크에 대한 현재 사용자 프로필의 결합 특징 표현과 연관된 벡터 특징을 원본 사용자 프로필의 결합 특징 표현과 연관된 벡터 특징과 비교하여 컴퓨팅 디바이스 접근을 위한 현재 사용자 프로필을 인증하는 것을 포함할 수 있다.

[185] 시스템(400)은 변경된 이산 에포크에 대한 결합 특징 표현을 생성하기 위해 합성된 음향 특징 및/또는 증강된 랜드마크 좌표와 연관된 하나 이상의 변경된 벡터 특징을 추출한다. 프로필 처리 요청은 원본 사용자 프로필의 초기 등록 또는 컴퓨팅 디바이스 접근을 위한 현재 에포크에 대한 사용자 프로필 인증을 포함할 수 있다. 즉, 시스템(400)은 사용자를 처음 등록할 때 프로필을 생성하고, 이후 사용자가 디바이스에 접근하려고 할 때마다 현재 에포크에 대한 사용자 프로필을 인증한다.

[186] 사용자 프로필 인증은 현재 에포크에 대한 현재 사용자 프로필의 결합 특징 표현과 연관된 벡터 특징을 원본 사용자 프로필의 결합 특징 표현과 연관된 벡터 특징과 비교하여 수행된다. 즉, 시스템(400)은 현재 사용자의 특징과 등록된 원본 사용자의 특징을 비교하여 일치하는지 확인한다. 만약 두 특징 벡터 사이의 유사도가 충분히 높으면 사용자 인증에 성공하고, 디바이스 접근이 허용된다.

[187] 일 실시예에 따르면, 시스템은 디바이스와 연관된 프로필 처리 요청에 응답하여 음향 신호를 방출한다. 이 과정에서 시스템은 사용자의 환경에 맞춰 특정 주파수의 음향 신호를 생성한다. 예를 들어, 사용자가 음악을 듣기 위해 음향 신호를 요청하면, 시스템은 고품질의 오디오 신호를 방출하여 음악의 명료성을 높인다.

[188] 사용자와 연관된 하나 이상의 깊이 부분의 고유한 윤곽으로부터 반사된 음향 신호를 기반으로 맞춤화된 일련의 에코 음향 신호를 수신한다. 예를 들어, 사용자가 특정 공간에서 음향 신호를 방출하면, 그 신호가 벽이나 가구에 반사되어 돌아오는 에코 음향 신호가 수집된다. 이 에코 신호는 사용자의 위치나 공간의 구조에 대한 정보를 포함하고 있어, 시스템은 이를 분석하여 환경에 대한 이해를 높인다.

[189] 시스템은 분류 모델을 훈련시키기 위해 에코 음향 신호와 연관된 하나 이상의 영역 세그먼트를 추출한다. 예를 들어, 시스템은 수신된 에코 신호에서 특정 주파수 대역의 변화를 분석하여 신호의 주요 세그먼트를 식별한다. 이를 통해, 각 세그먼트는 사용자의 환경이나 위치에 대한 중요한 정보를 담고 있어, 이러한 세그먼트를 기반으로 분류 모델을 생성할 수 있다.

[190] 일 실시예에 따르면, 시스템은 분류 모델을 기반으로 결합 특징 표현을 추출한다. 이 과정에서 시스템은 사용자로부터 수집된 음향 특징과 시각적 랜드마크 특징을 동시에 분석하여 결합 특징 표현을 생성한다. 예를 들어, 사용자가 특정 공간에서 말할 때, 그 음성과 함께 해당 공간의 영상 정보(예: 카메라로 촬영된 이미지)가 함께 수집된다.

[191] 결합 특징 표현의 예측에 사용되는 벡터 기반 분류 모델을 생성한다. 이 모델은 음향 데이터와 시각적 데이터를 결합하여 분석할 수 있도록 설계된다. 예를 들어, 시스템은 음성의 주파수 성분과 함께 사용자의 위치를 시각적으로 표시하는 랜드마크 정보를 결합하여 모델을 훈련시킨다.

[192] 결합 특징 표현의 예측을 기반으로 결합 특징 표현이 에코-시그니처와 연관되어 있는지 여부를 결정한다. 예를 들어, 시스템은 특정 음향 패턴이 특정 환경에서 발생할 가능성을 평가하여, 이 패턴이 이전에 수집된 에코-시그니처와 유사한지를 판단한다.

[193] 사용자 프로필과 연관된 결합 특징 표현에는 이산 에포크에 대한 하나 이상의 깊이 부분의 고유한 윤곽과 관련된 추출된 랜드마크 좌표 및 음향 특징이 포함된다. 예를 들어, 시스템은 사용자의 신체적인 형태나 특징을 기반으로 한 윤곽을 분석하여, 해당 윤곽에 대한 랜드마크 좌표를 생성한다.

[194] 사용자 프로필과 연관된 결합 특징 표현을 생성할 때, 변경된 이산 에포크에 대한 사용자와 연관된 하나 이상의 깊이 부분의 고유한 윤곽과 관련된 합성된 음향 특징 및 증강된 랜드마크 좌표로 결합 특징 표현을 보강할 수 있다. 예를 들어, 새로운 환경에서 사용자가 생성한 음향 특징을 시스템이 학습하여, 이를 기존의 프로필 데이터와 결합하여 보다 정교한 사용자 모델을 생성할 수 있다.

[195] 일 실시예에 따르면, 시스템은 변경된 이산 에포크에 대한 결합 특징 표현을 생성하기 위해 합성된 음향 특징 및/또는 증강된 랜드마크 좌표와 연관된 하나 이상의 변경된 벡터 특징을 추출한다. 예를 들어, 사용자가 새로운 환경에서 음성을 생성할 때, 시스템은 그 환경에 특화된 음향 특징을 수집하고 이를 기반으로 벡터 특징을 업데이트한다.

[196] 컴퓨팅 디바이스와 연관된 프로필 처리 요청은 원본 사용자 프로필의 초기 등록 또는 컴퓨팅 디바이스 접근을 위한 현재 에포크에 대한 사용자 프로필 인증을 포함할 수 있다. 예를 들어, 사용자가 새로운 기기를 사용할 때, 시스템은 원본 프로필 정보를 확인하여 접근 권한을 부여하는 과정을 진행한다.

[197] 컴퓨팅 디바이스와 연관된 프로필 처리 요청은 또한 현재 에포크에 대한 현재 사용자 프로필의 결합 특징 표현과 연관된 벡터 특징을 원본 사용자 프로필의 결합 특징 표현과 연관된 벡터 특징과 비교하여 컴퓨팅 디바이스 접근을 위한 현재 사용자 프로필을 인증하는 것을 포함할 수 있다. 예를 들어, 사용자가 특정 시간에 특정 장소에서 생성한 음향 패턴을 기반으로, 시스템은 이 패턴이 원본 프로필의 데이터와 일치하는지를 확인하여 사용자의 신원을 인증한다. 이러한 과정을 통해 시스템은 보다 안전하고 신뢰성 있는 사용자 인증을 수행할 수 있다.

[198] 이상에서 설명된 실시예들은 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치, 방법 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPGA(field programmable gate array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다.

[199]

[200] 본 발명의 일 실시예에 따른 프로세서는 컴퓨팅 자원의 효율적 사용을 위해 '동적 리소스 스위칭(Dynamic Resource Switching)'을 수행한다. 평상시에는 연산 부하가 적은 통계 기반 모델(제3 AI 모델)만을 활성화하여 1차 스크리닝을 수행하고, 1차 딥보이스 확률이 임계값을 초과하는 의심 구간이 탐지된 경우에만 대용량 메모리를 점유하는 딥러닝 모델(제1, 제2 AI 모델)을 메모리에 로드하여 정밀 분석을 수행한다. 이를 통해 상시 모니터링 환경에서의 CPU 점유율과 전력 소모를 최소화하며, 실시간 탐지 속도(Latency)를 보장하는 하드웨어 최적화가 적용된다.

[201]

[202] 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소 (processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[203] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(signal wave)에 영구적으로, 또는 일시적으로 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록매체에 저장될 수 있다.

[204] 이상과 같이 실시예들이 비록 한정된 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기를 기초로 다양한 기술적 수정 및 변형을 적용할 수 있다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

[205] [산업상 이용가능]

[206] 본 발명은 인공지능 기반의 음성 분석 기술을 필요로 하는 정보 통신 보안 산업, 보이스피싱 예방 시스템 및 금융 보안 서비스 분야에서 이용 가능하다.

[207] 또는, 본 발명은 딥보이스 탐지 및 사용자 신원 확인이 요구되는 다양한 디지털 콘텐츠 보안 산업 및 통신 서비스 분야에서 산업상 이용가능성이 있다.

청구범위

[청구항 1] 생성형 인공지능에 의해 생성된 딥보이스를 검출하는 시스템에 있어서,
인스트럭션들을 저장하는 메모리; 및 상기 인스트럭션들을 실행하는 하나 이상의 프로세서를 포함하고,
상기 프로세서는, 입력된 음성 데이터에서 주파수 특성 및 시간적 변화 패턴을 포함하는 특징 벡터를 추출하고, 상기 추출된 특징 벡터를 음향 모델인 제1 AI 모델, 언어 모델인 제2 AI 모델, 및 패턴 인식 모델인 제3 AI 모델을 포함하는 앙상블 인공지능 모델에 입력하여 각각의 개별 딥보이스 확률을 산출하며, 상기 입력된 음성 데이터의 신호 품질(Signal Quality)을 측정하여 신호 대 잡음비(SNR)를 산출하고, 상기 산출된 SNR 값에 기초하여 상기 각 AI 모델에 적용될 가중치를 실시간으로 변동시키는 적응형 가중치 할당(Adaptive Weight Allocation) 알고리즘을 통해 최종 딥보이스 확률을 연산하고, 상기 최종 딥보이스 확률이 임계값을 초과하는 경우, 딥보이스 의심 경보를 생성하도록 구성된, 딥보이스 검출 및 사고 예방 인공지능 솔루션 제공 시스템.

[청구항 2] 제1항에 있어서,
상기 프로세서는, 컴퓨팅 자원의 부하를 제어하기 위하여 동적 리소스 스위칭(Dynamic Resource Switching)을 수행하되, 상기 입력된 음성 데이터를 연산량이 기설정된 기준 미만인 상기 제3 AI 모델을 이용하여 1차 분석하고, 상기 1차 분석 결과 딥보이스 확률이 예비 임계값을 초과하는 구간이 검출된 경우에 한하여, 상기 메모리의 활성 영역에 딥러닝 기반의 상기 제1 AI 모델 및 제2 AI 모델을 로드(Load)하여 정밀 분석을 수행하도록 구성된,
딥보이스 검출 및 사고 예방 인공지능 솔루션 제공 시스템.

[청구항 3] 제1항에 있어서,
상기 프로세서는, 상기 최종 딥보이스 확률이 상기 임계값을 초과하는 경우, 사용자 단말을 통해 발신자의 신원을 확인하기 위한 추가 질문을 제시하고, 상기 추가 질문에 대한 발신자의 응답 음성 신호를 수신하여, 질의 시점부터 응답 시점까지의 지연 시간(Latency) 및 응답 음성 파형의 지터(Jitter)와 쉬머(Shimmer) 변동률을 측정하며, 상기 지연 시간이 기설정된 인간 반응 범위를 벗어나거나, 상기 지터 변동률이 기설정된 인공 합성음 패턴 범위에 포함되는 경우, 상기 응답의 내용과 무관하게 상기 최종 딥보이스 확률에 페널티 가중치를 가산하여 업데이트하는 생체 활성 감지(Liveness Detection) 피드백을 수행하도록 구성된,
딥보이스 검출 및 사고 예방 인공지능 솔루션 제공 시스템.

[청구항 4] 제1항에 있어서,

상기 적응형 가중치 할당 알고리즘은, 상기 SNR 값이 기준치 미만인 저품질 신호일 경우, 잡음에 민감한 상기 제1 AI 모델(음향)의 가중치를 감소시키고 문맥 의존적인 상기 제2 AI 모델(언어)의 가중치를 증가시키며,
상기 최종 딥보이스 확률(P_{final})을 다음의 수식에 기초하여 산출하며, 여기서, w_i는 SNR에 따라 변동되는 각 모델의 가중치, p_i는 각 모델의 출력 확률인,

$$P_{final} = \frac{\sum(w_i \times p_i)}{\sum w_i}$$

딥보이스 검출 및 사고 예방 인공지능 솔루션 제공 시스템.

[청구항 5] 제1항에 있어서,
상기 제1 AI 모델은 멜 스펙트로그램(Mel-Spectrogram)을 입력받는 CNN(Convolutional Neural Network) 기반 모델이고, 상기 제2 AI 모델은 텍스트로 변환된 발화 내용을 입력받는 트랜스포머(Transformer) 기반 언어 모델이며, 상기 제3 AI 모델은 음성의 피치(Pitch) 및 에너지(Energy) 변화 통계치를 입력받는 패턴 인식 모델인,
딥보이스 검출 및 사고 예방 인공지능 솔루션 제공 시스템.

[청구항 6] 컴퓨팅 장치에 의해 수행되는 딥보이스 검출 방법에 있어서,
입력된 음성 데이터에서 주파수 특성 및 시간적 변화 패턴을 포함하는 특징 벡터를 추출하는 단계;
상기 추출된 특징 벡터를 음향 모델인 제1 AI 모델, 언어 모델인 제2 AI 모델, 및 패턴 인식 모델인 제3 AI 모델을 포함하는 앙상블 모델에 입력하여 개별 확률을 산출하는 단계;
상기 입력된 음성 데이터의 신호 품질을 분석하여 신호 대 잡음비(SNR)를 측정하는 단계;
상기 측정된 SNR 값에 따라 상기 복수의 AI 모델 각각에 적용할 가중치를 동적으로 결정하는 적응형 가중치 할당을 수행하여 최종 딥보이스 확률을 연산하는 단계; 및
상기 최종 딥보이스 확률이 임계값을 초과하는 경우, 사용자 단말에 경고 메시지를 표시하거나 수사 기관 서버로 신고 신호를 전송하는 단계를 포함하는,
딥보이스 검출 및 사고 예방 인공지능 솔루션 제공 방법.

[청구항 7] 제6항의 방법을 컴퓨터에서 실행시키기 위한 프로그램을 기록한 컴퓨터 판독 가능 기록 매체.