METHOD, APPARATUS, AND SYSTEM FOR PROVIDING ARTIFICIAL INTELLIGENCE SOLUTION FOR DETECTING DEEP VOICE GENERATED BY GENERATIVE ARTIFICIAL INTELLIGENCE AND PREVENTING RELATED ACCIDENT

[Technical Field]

The present invention relates to a method, an apparatus, and a system for providing an artificial intelligence solution for detecting a deep voice generated using generative artificial intelligence (Generative AI) technology and preventing a related accident.

[Background Art]

Recently, with the development of artificial intelligence technology, deep voice technology that very sophisticatedly imitates a human voice has emerged. Such deep voice technology may be utilized in various fields such as speech synthesis and entertainment, but when used for malicious purposes, it may cause serious social problems such as voice phishing, impersonation, and defamation.

Conventional deep voice detection technology mainly uses a method of analyzing the characteristics of a voice signal. For example, features appearing in a deep voice are extracted through Mel-Frequency Cepstral Coefficient (MFCC) analysis, Linear Predictive Coding (LPC) analysis, or the like, and whether the voice is a deep voice is determined based thereon. However, these methods show limitations in effectively detecting deep voices that are becoming increasingly sophisticated due to the development of generative AI technology.

[Problem to be Solved]

The present invention aims to effectively detect a deep voice that has become more sophisticated due to the development of generative AI technology and to provide a solution for preventing a related accident.

[Means for Solving the Problem]

According to one aspect of the present invention, there is provided a system for providing an artificial intelligence solution for detecting a deep voice and preventing an accident, the system comprising a system for detecting a deep voice generated by generative artificial intelligence, comprising: a memory storing instructions; and one or more processors executing the instructions, wherein the processor is configured to: extract a feature vector including a frequency characteristic and a temporal change pattern from input voice data; input the extracted feature vector into an ensemble artificial intelligence model

including a first AI model that is an acoustic model, a second AI model that is a language model, and a third AI model that is a pattern recognition model, to calculate a respective individual deep voice probability; measure a signal quality of the input voice data to calculate a signal-to-noise ratio (SNR); calculate a final deep voice probability through an adaptive weight allocation algorithm that varies, in real time, a weight to be applied to each of the AI models based on the calculated SNR value; and generate a deep voice suspicion alert when the final deep voice probability exceeds a threshold value.

According to an embodiment of the present invention, the processor may be configured to perform dynamic resource switching to control a load of computing resources, wherein the processor performs a first analysis of the input voice data using the third AI model, the third AI model having a computational amount less than a preset criterion, and, only when a section in which the deep voice probability exceeds a preliminary threshold value is detected as a result of the first analysis, loads the deep learning-based first AI model and second AI model into an active area of the memory to perform a precise analysis.

According to an embodiment of the present invention, when the final deep voice probability exceeds the threshold value, the processor may be configured to perform liveness detection feedback by presenting an additional question for

verifying an identity of a caller through a user terminal, receiving a response voice signal of the caller to the additional question, measuring a latency from a point in time of the question to a point in time of the response and a jitter and shimmer variation rate of a response voice waveform, and, when the latency deviates from a preset human response range or the jitter variation rate falls within a preset artificial synthesized sound pattern range, adding a penalty weight to the final deep voice probability to update the final deep voice probability, regardless of the content of the response.

According to an embodiment of the present invention, in a case where the SNR value is a low-quality signal that is less than a reference value, the adaptive weight allocation algorithm may decrease a weight of the first AI model (acoustic), which is sensitive to noise, and increase a weight of the second AI model (language), which is context-dependent, and calculate the final deep voice probability (Pfinal) based on the following equation, wherein wi is a weight of each model that varies according to the SNR, and pi is an output probability of each model.

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

According to an embodiment of the present invention, the first AI model may be a Convolutional Neural Network (CNN)-based model that receives a Mel-

Spectrogram as input, the second AI model may be a Transformer-based language model that receives, as input, speech content converted into text, and the third AI model may be a pattern recognition model that receives, as input, pitch and energy change statistics of a voice.

According to another aspect of the present invention, there is provided a method for providing an artificial intelligence solution for detecting a deep voice and preventing an accident, the method being performed by a computing device and comprising: extracting a feature vector including a frequency characteristic and a temporal change pattern from input voice data; inputting the extracted feature vector into an ensemble model including a first AI model that is an acoustic model, a second AI model that is a language model, and a third AI model that is a pattern recognition model, to calculate an individual probability; measuring a signal-to-noise ratio (SNR) by analyzing a signal quality of the input voice data; calculating a final deep voice probability by performing adaptive weight allocation that dynamically determines a weight to be applied to each of the plurality of AI models according to the measured SNR value; and, when the final deep voice probability exceeds a threshold value, displaying a warning message on a user terminal or transmitting a report signal to an investigative agency server.

According to still another aspect of the present invention, there is provided a computer-readable recording medium having recorded thereon a program for executing the method on a computer.

[Effects of the Invention]

Using a generative AI model, a deep voice can be detected more accurately than by conventional methods. By training various generative AI models, various types of deep voices can be effectively detected.

Deep voices occurring in real time, such as during voice calls or online meetings, can be immediately detected to prevent damage. In addition, crimes such as voice phishing and impersonation using deep voices can be prevented.

[Brief Description of the Drawings]

FIG. 1 is a diagram illustrating a system for providing an artificial intelligence solution for detecting a deep voice generated by generative artificial intelligence and preventing a related accident according to an embodiment.

FIG. 2 is a diagram illustrating training of a neural network according to an embodiment.

FIG. 3 is a diagram illustrating a configuration of an artificial intelligence model according to an embodiment.

FIG. 4 is a block diagram illustrating a configuration of a system for providing an artificial intelligence solution for detecting a deep voice generated by

generative artificial intelligence and preventing a related accident according to an embodiment.

FIG. 5 is a flowchart illustrating a method for providing an artificial intelligence solution for detecting a deep voice generated by generative artificial intelligence and preventing a related accident according to an embodiment.

FIG. 6 is a flow diagram illustrating a method for providing an artificial intelligence solution for detecting a deep voice generated by generative artificial intelligence and preventing a related accident according to an embodiment.

FIG. 7 is a flow diagram illustrating a method for providing an artificial intelligence solution for detecting a deep voice generated by generative artificial intelligence and preventing a related accident according to an embodiment.

FIG. 8 illustrates an adaptive weight allocation process according to an embodiment of the present invention.

[Detailed Description]

Hereinafter, embodiments will be described in detail with reference to the accompanying drawings.

However, various changes may be made to the embodiments, and the scope of the patent application is not restricted or limited by these embodiments. It should be understood that all changes, equivalents, or substitutes with respect to the embodiments are included in the scope of the rights.

The specific structural or functional descriptions of the embodiments are disclosed only for illustrative purposes, and may be changed and implemented in various forms. Accordingly, the embodiments are not limited to a specific disclosed form, and the scope of the present specification includes changes, equivalents, or substitutes included in the technical spirit.

Terms such as first or second may be used to describe various components, but these terms should be interpreted only for the purpose of distinguishing one component from another component. For example, a first component may be referred to as a second component, and similarly, a second component may also be referred to as a first component.

When a component is referred to as being "connected" to another component, it should be understood that the component may be directly connected or coupled to the other component, but another component may exist therebetween.

Terms used in the embodiments are used only for the purpose of description and are not intended to be limiting. A singular expression includes a plural expression unless the context clearly indicates otherwise. In this specification, terms such as "comprise" or "have" are intended to designate that a feature, a number, a step, an operation, a component, a part, or a combination thereof described in the specification exists, and should be understood as not precluding in advance the possibility of the existence or addition of one or more

other features, numbers, steps, operations, components, parts, or combinations thereof.

Unless otherwise defined, all terms used herein, including technical or scientific terms, have the same meaning as commonly understood by a person having ordinary skill in the art to which the embodiments belong. Terms such as those defined in a commonly used dictionary should be interpreted as having a meaning consistent with the meaning they have in the context of the relevant technology, and unless explicitly defined in the present application, should not be interpreted in an ideal or excessively formal sense.

In addition, in describing with reference to the accompanying drawings, identical components are given identical reference numerals regardless of reference numerals in the drawings, and redundant descriptions thereof are omitted. In describing the embodiments, when it is determined that a detailed description of a related known technology may unnecessarily obscure the gist of the embodiments, the detailed description thereof is omitted.

The embodiments may be implemented as various types of products, such as a personal computer, a laptop computer, a tablet computer, a smartphone, a television, a smart home appliance, an intelligent vehicle, a kiosk, and a wearable device.

An artificial intelligence (AI) system is a computer system that implements human-level intelligence, and, unlike a conventional rule-based smart system, is a system in which a machine learns and makes decisions on its own. As an artificial intelligence system is used, its recognition rate improves and it is able to more accurately understand user preferences, and accordingly, conventional rule-based smart systems are gradually being replaced with deep learning-based artificial intelligence systems.

Artificial intelligence technology is composed of machine learning and element technologies utilizing machine learning. Machine learning is an algorithm technology that classifies/learns the features of input data on its own, and the element technologies are technologies that utilize machine learning algorithms such as deep learning to mimic functions of a human brain, such as cognition and judgment, and are composed of technical fields such as linguistic understanding, visual understanding, inference/prediction, knowledge representation, and motion control.

Various fields to which artificial intelligence technology is applied are as follows.

Linguistic understanding is a technology for recognizing and applying/processing human language/characters, and includes natural language processing, machine translation, a dialogue system, question answering, and speech recognition/synthesis. Visual understanding is a technology for

recognizing and processing an object like human vision, and includes object recognition, object tracking, image search, human recognition, scene understanding, spatial understanding, and image enhancement. Inference/prediction is a technology for judging information and logically inferring and predicting, and includes knowledge/probability-based inference, optimization prediction, preference-based planning, and recommendation. Knowledge representation is a technology for automating human experience information into knowledge data, and includes knowledge construction (data generation/classification) and knowledge management (data utilization). Motion control is a technology for controlling autonomous driving of a vehicle and movement of a robot, and includes movement control (navigation, collision, driving) and operation control (behavior control).

In general, in order to apply a machine learning algorithm to real life, learning is performed in a trial-and-error manner due to the nature of the basic methodology of machine learning. In particular, deep learning requires hundreds of thousands of repeated executions. Since it is impossible to execute this in an actual physical external environment, learning is instead performed through simulation by virtually implementing the actual physical external environment on a computer.

In the present invention, artificial intelligence (AI) refers to a technology that imitates human learning ability, reasoning ability, perception ability, and the like, and implements the same on a computer, and may include concepts such as machine learning and symbolic logic. Machine learning (ML) is an algorithm technology that classifies or learns the features of input data on its own. The technology of artificial intelligence, as an algorithm of machine learning, uses input data to

analyze the data, learn a result of the analysis, and make a determination or prediction based on a result of the learning. In addition, technologies that utilize machine learning algorithms to mimic functions of the human brain, such as cognition and judgment, may also be understood as belonging to the category of artificial intelligence. For example, technical fields such as linguistic understanding, visual understanding, inference/prediction, knowledge representation, and motion control may be included.

An artificial intelligence learning model or a neural network model may be designed to implement a human brain structure on a computer, and may include a plurality of network nodes having weights that simulate neurons of a human neural network. The plurality of network nodes may have a connection relationship with each other by simulating synaptic activity of neurons in which neurons exchange signals through a synapse. In the artificial intelligence learning model,

the plurality of network nodes may be located at layers of different depths and exchange data according to a convolutional connection relationship. The artificial intelligence learning model may be, for example, an artificial neural network model, a convolutional neural network (CNN) model, or the like. As an embodiment, the artificial intelligence learning model may be machine-learned according to a method such as supervised learning, unsupervised learning, or reinforcement learning. A machine learning algorithm for performing machine learning may use a decision tree, a Bayesian network, a support vector machine, an artificial neural network, Ada-boost, a perceptron, genetic programming, clustering, or the like.

Among these, a CNN is a type of multilayer perceptron designed to use minimal preprocessing. A CNN is composed of one or more convolutional layers and general artificial neural network layers stacked thereon, and additionally utilizes weights and pooling layers. Due to this structure, a CNN can fully utilize input data having a two-dimensional structure. Compared to other deep learning structures, a CNN shows good performance in both the image and speech fields. A CNN can also be trained through standard backpropagation. A CNN has the advantage of being easier to train than other feedforward artificial neural network techniques and using a small number of parameters.

A convolutional network is a neural network including sets of nodes having tied parameters. Increases in the size of available training data and the availability of computing power, combined with algorithmic advances such as piecewise linear units and dropout training, have greatly improved many computer vision tasks. In enormous datasets, such as the datasets available for many tasks today, overfitting is not significant, and increasing the size of the network improves test accuracy. Optimal use of computing resources becomes a limiting factor. To this end, distributed, scalable implementations of deep neural networks may be used.

FIG. 1 is a diagram illustrating a system for providing an artificial intelligence solution for detecting a deep voice generated by generative artificial intelligence and preventing a related accident according to an embodiment.

As illustrated in FIG. 1, a system (100) for providing an artificial intelligence solution for detecting a deep voice generated by generative artificial intelligence and preventing a related accident may include a plurality of user terminals (110-1, ... 110-n), a server (120), and a database (130). According to an embodiment, although the database (130) is illustrated as being configured separately from the server (120), it is not limited thereto, and the database (130) may be provided within the server (120). For example, the server (120) may include a plurality of artificial intelligences for performing a

machine learning algorithm. According to an embodiment, the plurality of user terminals (110-1, ... 110-n), the server (120), and the database (130) may be communicatively connected to each other through a network (N).

The network (N) may enable wireless or wired communication among the plurality of user terminals (110-1, ... 110-n), the server (120), the database (130), and the like. For example, the network may perform wireless communication according to a method such as long-term evolution (LTE), LTE Advanced (LTE-A), code division multiple access (CDMA), wideband CDMA (WCDMA), Wireless BroadBand (WiBro), wireless fidelity (WiFi), Bluetooth, near field communication (NFC), Global Positioning System (GPS), or global navigation satellite system (GNSS). For example, the network (N) may also perform wired communication according to a method such as universal serial bus (USB), high definition multimedia interface (HDMI), recommended standard 232 (RS 232), or plain old telephone service (POTS).

The database (130) may store various data. Data stored in the database (130) is data that is acquired, processed, or used by at least one component of the plurality of user terminals (110-1, ... 110-n) and the server (120), and may include software (e.g., a program). The database (130) may include a volatile and/or non-volatile memory.

FIG. 2 is a diagram illustrating training of a neural network according to an embodiment.

As illustrated in FIG. 2, a training device may train a neural network (123) to categorize review responses received from the plurality of user terminals (110-1, ... 110-n) by item. In addition, the training device may train the neural network (123) to extract a user's stay history from the user's movement path information. According to an embodiment, the training device may be a separate entity from the server (120), but is not limited thereto.

The neural network (123) includes an input layer (121) to which training samples are input and an output layer (125) that outputs training outputs, and may be trained based on a difference between the training outputs and labels. Here, the labels may be defined based on items corresponding to the review responses, and may be defined based on a user's stay history corresponding to the movement path information. The neural network (123) is connected as a group of a plurality of nodes, and is defined by weights between the connected nodes and an activation function that activates the nodes.

The training device may train the neural network (123) using a gradient descent (GD) technique or a stochastic gradient descent (SGD) technique. The training device may use a loss function designed based on outputs and labels of the neural network.

The training device may calculate a training error using a predefined loss function. The loss function may be predefined with a label, an output, and a parameter as input variables, where the parameter may be set by weights in the neural network (123). For example, the loss function may be designed in the form of mean square error (MSE), entropy, or the like, and various techniques or methods may be employed in an embodiment in which the loss function is designed.

The training device may find weights that affect the training error using a backpropagation technique. Here, the weights are relationships between nodes in the neural network (123). The training device may use an SGD technique using labels and outputs to optimize the weights found through the backpropagation technique. For example, the training device may update weights of a loss function defined based on labels, outputs, and weights using an SGD technique.

According to an embodiment, the training device may extract first objects from a review response, obtain first labels that are items corresponding to the first objects, apply the first objects to a first neural network to generate first training outputs corresponding to the first objects, and train the first neural network based on the first training outputs and the first labels.

The training device may extract second objects from movement path information, obtain second labels that are a user's stay history corresponding to the second objects, apply the second objects to a second neural network to generate second training outputs corresponding to the second objects, and train the second neural network based on the second training outputs and the second labels.

According to an embodiment, the training device may generate first training feature vectors based on compositional features, positional features, and pattern features of a review response.

Various methods may be employed to extract a feature.

According to an embodiment, the training device may generate second training feature vectors based on compositional features, length features, and pattern features of movement path information. Various methods may be employed to extract a feature.

According to an embodiment, the training device may apply the first training feature vectors to the neural network (123) to obtain training outputs. The training device may train a review item extraction algorithm of the neural network (123) based on the training outputs and the first labels. The training device may calculate training errors corresponding to the training outputs, and may train the review item extraction algorithm of the neural network (123) by

optimizing a connection relationship of nodes in the neural network (123) to minimize the training errors. The server (120) may extract items from a review response using the trained first neural network.

According to an embodiment, the training device may apply the second training feature vectors to the neural network (123) to obtain training outputs. The training device may train a user-stay-history acquisition algorithm of the neural network (123) based on the training outputs and the second labels. The training device may calculate training errors corresponding to the training outputs, and may train the user-stay-history acquisition algorithm of the neural network (123) by optimizing a connection relationship of nodes in the neural network (123) to minimize the training errors. The server (120) may obtain a user's stay history from movement path information using the trained second neural network.

FIG. 3 is a diagram illustrating a configuration of an artificial intelligence model according to an embodiment.

An artificial intelligence model according to an embodiment may include an input layer, a hidden layer, and an output layer.

The input layer is a layer related to an input value input to the artificial intelligence model.

In the hidden layer, a feature map may be output by performing a multiply-accumulate (MAC) operation and an activation operation on an input value.

The MAC operation may be an operation of multiplying an input value by a corresponding weight, respectively, and summing the multiplied values.

The activation operation may be an operation of inputting a result of the MAC operation into an activation function and outputting a result value. The activation function may be of various types. For example, the activation function may include, but is not limited to, a sigmoid function, a tangent function, a ReLU function, a leaky ReLU function, a maxout function, and/or an ELU function.

The hidden layer may be composed of at least one layer. For example, when the hidden layer is composed of a first hidden layer and a second hidden layer, the first hidden layer may perform a MAC operation and an activation operation based on an input value of the input layer to output a feature map, and the feature map, which is a result value of the first hidden layer, may become an input value of the second hidden layer. The second hidden layer may perform a MAC operation and an activation operation based on the feature map, which is a result value of the first hidden layer.

The output layer may be a layer related to a result value of an operation performed in the hidden layer.

In one embodiment, the training model may learn syllable (character) patterns frequently combined and used in a given corpus to automatically learn boundaries of compound words and named entities, integrate object information of a first UI source and object information rendered in a browser to generate an object information file for training, generate training data for training a deep learning network using the object information file for training, receive data of various domains of a support system, standardize the data of the various domains into an integrated format based on at least one standardization method corresponding to each of the various domains, learn and infer data of a specific domain, determine, from the data of the specific domain, information to be delivered for the standardization, and perform post-processing on data from the various domains. The first UI source includes an XML file, the object information file for training includes an input JSON file for learning features and an output JSON file that is answer (label) data during training, the output JSON file includes a file containing DOM tree information of HTML implemented in compliance with web standards, the various domains include at least one of a radio access network (RAN), a transport, or a core, and the post-processing may include a correlation function.

FIG. 4 is a block diagram illustrating a configuration of a system for providing an artificial intelligence solution for detecting a deep voice generated by

generative artificial intelligence and preventing a related accident according to an embodiment.

A system (400) according to an embodiment may include a processor (420) and a memory (430), and some of the illustrated components may be omitted or replaced. The system (400) according to an embodiment may be a server or a terminal. According to an embodiment, the processor (420) is a component capable of performing an operation or data processing related to control and/or communication of each component of the system (400), and may be composed of one or more processors. The memory (430) may store information related to the method described above or a program in which the method described above is implemented. The memory (430) may be a volatile memory or a non-volatile memory. The memory (430) may store various file data, and the stored file data may be updated according to an operation of the processor (420).

According to an embodiment, the processor (420) may execute a program and control the device (400). Code of the program executed by the processor (420) may be stored in the memory (430). Operations of the processor (420) may be performed by loading instructions stored in the memory (430). The system (400) is connected to an external device (e.g., a personal computer or a network) through an input/output device (not shown) and may exchange data.

The system (400) according to an embodiment includes a processor (420) and a memory (430). The system (400) according to an embodiment may be the above-described server or terminal. The processor (420) may include at least one device described above with reference to FIGS. 1 to 3, or may perform at least one method described above with reference to FIGS. 1 to 3. The memory (430) may store information related to the method described above or a program in which the method described above is implemented. The memory (430) may be a volatile memory or a non-volatile memory.

The system (400) according to an embodiment includes a processor (420) and a memory (430). The system (400) according to an embodiment may be the above-described server or terminal. The processor (420) may include at least one device described above with reference to FIGS. 1 to 3, or may perform at least one method described above with reference to FIGS. 1 to 3. The memory (430) may store information related to the method described above or a program in which the method described above is implemented. The memory (430) may be a volatile memory or a non-volatile memory.

The final deep voice probability (Pfinal) is calculated by applying weights (w1, w2, w3) to respective output probabilities (p1, p2, p3) of the first AI model (acoustic), the second AI model (language), and the third AI model (pattern).

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

At this time, each weight wi is not a fixed value, but varies in real time based on a signal-to-noise ratio (SNR) of the input voice signal and a confidence score of each model. For example, when the background noise of the input voice is severe such that the SNR is less than a reference value (e.g., 10 dB), an adaptive weight allocation algorithm is applied that decreases the weight w1 of the first AI model (acoustic), which is vulnerable to noise, by a factor of 0.5 and increases the weight w2 of the context-dependent second AI model (language) by a factor of 1.5.

FIG. 5 is a flowchart illustrating a method for providing an artificial intelligence solution for detecting a deep voice generated by generative artificial intelligence and preventing a related accident according to an embodiment.

Although process steps, method steps, algorithms, and the like are described in a sequential order in the flowchart of FIG. 5, such processes, methods, and algorithms may be configured to operate in any suitable order. In other words, steps of the processes, methods, and algorithms described in various embodiments of the present invention do not need to be performed in the order described herein. In addition, although some steps are described as being performed asynchronously, in another embodiment, some of these steps may

be performed simultaneously. In addition, the illustration of a process by depiction in a drawing does not mean that the illustrated process excludes other variations and modifications thereto, does not mean that any of the illustrated process or steps thereof is essential to one or more of various embodiments of the present invention, and does not mean that the illustrated process is preferred.

In operation S510, a system (e.g., the system (400) of FIG. 4) for detecting a deep voice created through generative artificial intelligence (AI) and preventing a related crime may extract a deep-voice-suspected feature under control of a processor (e.g., the processor (420) of FIG. 4). The system (400) may receive voice data (e.g., a phone call) in real time, and may extract the deep-voice-suspected feature by analyzing a frequency characteristic, a temporal change pattern, and fine noise in the input voice data.

For example, when contents of a phone call are input to the system, the system collects the voice data in real time. When an abnormal pattern is found in a specific frequency band by analyzing the frequency characteristic of the input voice data, this may be a deep-voice-suspected feature. For example, when a user says "Hello, this is Kim Cheol-su," the system analyzes a frequency pattern of this voice to identify a difference from a normal speaker's voice. In addition, the system detects a change in the speech rate or intonation of the

voice by analyzing the temporal change pattern. If the user speaks faster than usual or uses an abnormal intonation, this is regarded as a deep-voice-suspected feature. Finally, the system evaluates the degree of background noise or distortion of the voice by analyzing fine noise. For example, if mechanical noise or distorted voice occurring during a call is detected, this serves as an additional suspicious factor.

It may be assumed that user A is on a call with voice phishing criminal B. In this case, criminal B may use a voice modulated into a female voice using deep voice technology. The system (400) receives the contents of the call between A and B in real time. The system (400) may detect that energy in a frequency band lower than that of a typical female voice appears high in B's voice. The system (400) detects that intervals between words or changes in intonation in B's voice are unnatural or show a constant pattern. The system (400) may detect, in B's voice, a specific noise pattern present in the deep learning model training data, or fine digital noise that does not generally appear in a human voice.

In operation S520, the system (400) may determine whether the voice is a deep voice. The system (400) may determine whether the voice is a deep voice by analyzing the extracted feature using a generative AI model.

The AI model is a model specialized for deep voice detection, and has high accuracy in determining whether a voice is a deep voice by learning from a large amount of deep voice data and actual human voice data. As a result of analyzing the input features, the AI model finds characteristic patterns occurring in the deep voice generation process in B's voice and determines it to be a deep voice.

According to an embodiment, the system (400) determines whether the voice is a deep voice. In this step, the feature extracted in the previous step is analyzed using a generative AI model. For example, the system compares and analyzes a frequency pattern and a temporal change pattern of the voice using a deep-learning-based acoustic model. If the system reaches a conclusion that "this voice is likely to have been generated by generative AI," the voice is determined to be a deep voice. For example, if an abnormal pattern repeatedly appears in a specific frequency band or the pronunciation lacks consistency, the system may recognize this as a deep voice. This determination process is performed in real time, and the system comprehensively analyzes the output results of each model to make a final decision.

In operation S530, the system (400) may display a warning message on a user terminal and automatically report it. When the voice is determined to be a deep voice, the system (400) may display a warning message reading "Caution! A

deep-voice-suspected voice has been detected." on the user terminal, and may control the system to automatically report information related to the deep-voice-suspected voice data to the Cyber Safety Bureau of the National Police Agency.

According to an embodiment, the system (400) may analyze an extracted feature using a generative artificial intelligence model (AI), calculate a deep voice probability based on a result of the analysis, and display a warning message when the deep voice probability exceeds a specified level and the voice is determined to be a deep voice.

When the voice is determined to be a deep voice, the system may prevent crime by displaying a warning message on the user terminal and automatically reporting it. For example, the system displays a warning message reading "Caution! A deep-voice-suspected voice has been detected." on the screen of user A's smartphone. This message notifies that the voice has been determined to be a deep voice, and helps the user immediately recognize the situation. In addition, the system controls to automatically report information related to the deep-voice-suspected voice data to the Cyber Safety Bureau of the National Police Agency. At this time, the reported information includes the caller's phone number, the call time, the deep voice probability, and part of the contents of the call. For example, the system makes a report in the form of "a deep-

voice-suspected voice occurring at 15:30 on October 17, 2024, from user A's phone number 010-1234-5678."

When the AI model determines B's voice to be a deep voice, the system (400) immediately displays a warning message reading "Caution! A deep-voice-suspected voice has been detected." on A's smartphone screen. At the same time, the system automatically reports related information, such as B's deep voice data, the call start time, and the call contents, to the Cyber Safety Bureau of the National Police Agency. Through this, A can be alert to voice phishing crime and prevent damage, and the police can track down the criminal and utilize the reported information for investigation.

FIG. 6 is a flow diagram illustrating a method for providing an artificial intelligence solution for detecting a deep voice generated by generative artificial intelligence and preventing a related accident according to an embodiment.

Although process steps, method steps, algorithms, and the like are described in a sequential order in the flowchart of FIG. 6, such processes, methods, and algorithms may be configured to operate in any suitable order. In other words, steps of the processes, methods, and algorithms described in various embodiments of the present invention do not need to be performed in the order described herein. In addition, although some steps are described as being performed asynchronously, in another embodiment, some of these steps may

be performed simultaneously. In addition, the illustration of a process by depiction in a drawing does not mean that the illustrated process excludes other variations and modifications thereto, does not mean that any of the illustrated process or steps thereof is essential to one or more of various embodiments of the present invention, and does not mean that the illustrated process is preferred.

In operation S610, a system (e.g., the system (400) of FIG. 4) may extract a frequency characteristic from input voice data under control of a processor (e.g., the processor (420) of FIG. 4). The system (400) may extract a frequency characteristic, including Mel-Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC) coefficients, and a spectrogram, from the input voice data, and may extract a temporal change pattern including a pitch change, a formant change, and an energy change.

First, the system (400) divides the voice data into small time units (e.g., 20 ms) and performs a Fourier transform on each unit to obtain a frequency spectrum. Mel-Frequency Cepstral Coefficients (MFCC) are extracted from this spectrum, which is a compression of frequency information by imitating the way the human auditory system perceives frequency. MFCC is a widely used feature in speech recognition, and has the advantage of being particularly robust to noise.

The system (400) then extracts Linear Predictive Coding (LPC) coefficients. LPC is a technique for efficiently representing the spectral envelope of a voice signal, and analyzes the difference by predicting a current voice sample as a linear combination of past samples. Through this, the resonance characteristics of the vocal tract can be identified, and the characteristics of the voice can be effectively represented.

A spectrogram is a visual representation of a change in frequency over time, and is useful for analyzing the energy distribution of a voice signal. The processor (420) analyzes the spectrogram to track a pitch change, that is, a change in the pitch of the voice. In addition, the processor (420) analyzes a formant change, where a formant is a resonant frequency of the vocal tract and is an important factor determining the characteristics of a vowel. Finally, the processor (420) analyzes an energy change over time to determine a change in the intensity of the voice.

In operation S620, the system (400) may calculate a final deep voice probability by weighted-averaging the output results of each model. The system (400) may extract fine noise including background noise, breathing sound, and lip sound, extract a pronunciation pattern including phoneme duration, resonance features, and articulation manner, extract a voice pause time including an interval between words and an interval between sentences, extract context

information including speech content, conversation topic, and situational information, input the extracted features into an ensemble AI model including a deep-learning-based acoustic model, a natural-language-processing-based language model, and a statistics-based pattern recognition model to determine whether the voice is a deep voice, and calculate a final deep voice probability by weighted-averaging the output results of each model.

First, the system (400) separates fine noise, such as background noise, breathing sound, and lip sound, from the voice data. Since such noise may interfere with deep voice determination, it is removed or attenuated for accurate analysis.

The system (400) analyzes a pronunciation pattern, such as phoneme duration, resonance features, and articulation manner. Since a deep voice may differ subtly from a human's pronunciation,

this difference in pattern is utilized to determine whether the voice is a deep voice.

The system (400) analyzes a voice pause time, that is, the interval between words and the interval between sentences. Since a deep voice, unlike a human, does not require breathing, the pause-time pattern may appear differently from that of a human.

The system (400) analyzes context information such as speech content, conversation topic, and situational information. A deep voice may involve speech that does not fit the context, or may fail to properly understand the

conversation topic. This context information is utilized to determine whether the voice is a deep voice. The extracted features are input into an ensemble AI model to determine whether the voice is a deep voice.

The ensemble AI model combines a deep-learning-based acoustic model, a natural-language-processing-based language model, and a statistics-based pattern recognition model.

The acoustic model determines whether the voice is a deep voice by analyzing acoustic features of the voice data. The language model determines whether the voice is a deep voice by analyzing grammatical and semantic errors of the speech content. The pattern recognition model determines whether the voice is a deep voice by analyzing statistical patterns of the voice data. Each model independently determines whether the voice is a deep voice, and the system (400) calculates the final deep voice probability by weighted-averaging the output results of each model. The weight is determined according to the performance of each model, and a higher weight is given to a model with better performance. Through this, the strengths of each model are combined and their weaknesses are compensated for, thereby increasing the accuracy of deep voice determination.

In operation S630, when the voice is determined to be a deep voice, the system (400) generates a warning and presents an additional question capable of

verifying the identity of the caller. Specifically, the system (400) determines the corresponding voice to be a deep voice when the calculated deep voice probability is equal to or greater than a preset threshold value (e.g., a user-set value or a system default value). When the voice is determined to be a deep voice, the system immediately displays, on the user terminal, a visual warning message reading "Caution! A deep-voice-suspected voice has been detected.", and simultaneously generates a vibration or a warning sound so that the user can immediately recognize the danger. Thereafter, the system (400) presents an additional question, such as "Does the caller know their own date of birth?" or "What was the topic of the recent conversation with the caller?", to verify whether the caller is an actual acquaintance. If the caller fails to give an accurate answer thereto (e.g., "It is January 1, 1990.") or hesitates, the system may utilize this as additional grounds for the deep voice determination.

In operation S640, the system (400) may reevaluate the deep voice probability by analyzing the collected user response. The system (400) may collect the user's response in the form of voice or text, reevaluate the deep voice probability by analyzing the collected user response, and provide the reevaluated deep voice probability to the user in the form of a numerical value or a graph, such as "deep voice probability: XX%."

The system reevaluates the deep voice probability by analyzing the collected user response. When the user responds in the form of voice or text, the system collects and analyzes this response. For example, when the user responds to the additional question with "I know the date of birth," the system processes this response to obtain information necessary for verifying the identity of the caller.

In this process, the system evaluates the consistency of the user's response. If the user responds "The recent conversation topic was travel," the system compares this with previous conversation records to determine whether this response is credible. Thereafter, the system reevaluates the deep voice probability based on the collected user response. For example, if the initial deep voice probability was 85%, this probability may be lowered to 50% due to the accurate information provided by the user. The reevaluated deep voice probability is provided to the user in the form of a numerical value or a graph, such as "deep voice probability: XX%." Through this, the user can clearly understand the current situation.

The user may respond, in the form of voice or text, to the additional question presented by the system (400). The system (400) collects the user's response and analyzes the content of the response using natural language processing technology. The deep voice probability is reevaluated based on the result of

the analysis. For example, if the user fails to accurately answer the question, or if the content of the answer is ambiguous or inconsistent, the deep voice probability is increased. Conversely, if the user answers the question accurately and clearly, the deep voice probability is decreased. The reevaluated deep voice probability is displayed to the user as a numerical value, such as "deep voice probability: XX%," or visually represented in the form of a graph, such as a bar graph or a pie chart. Through this, the user can more intuitively determine whether the voice is a deep voice.

In addition to simply whether the answer matches, the system analyzes a 'liveness response' in the received response voice signal of the user and reflects it in the probability reevaluation. Specifically, the system measures, in milliseconds (ms), a latency from a point in time of presenting the question to a point in time of starting the response, and calculates a jitter and shimmer variation rate of the response voice waveform. If the measured response latency deviates from a typical human response speed range (e.g., 200 to 500 ms), or if the jitter variation rate of the waveform shows a flat pattern characteristic of an artificial synthesized sound (e.g., less than 0.5%), the system performs a feedback loop that adds a penalty weight (e.g., + 0.2) to the deep voice probability and updates the final probability upward, even if the content of the answer is correct.

In operation S650, the system (400) may periodically update the deep voice detection model. The system (400) may, according to the user's selection, immediately block a deep-voice-suspected call, may, according to the user's selection, record the contents of the call and store it in the user's cloud storage, may analyze the recorded file and utilize it to determine whether the voice is a deep voice, and may periodically update the deep voice detection model. The update may include retraining the model using new data reflecting the latest deep voice generation technology. The system (400) may automatically deploy the updated model to the user terminal, and may control the system to encrypt information related to the deep-voice-suspected voice data, including at least one of a caller number, a call time, a deep voice probability, a user response, and a recorded file, and automatically report it to the Cyber Safety Bureau of the National Police Agency.

The system periodically updates the deep voice detection model. When the user selects, in a specific situation, to immediately block a deep-voice-suspected call, the system blocks the corresponding call. For example, when the user determines that the caller's voice is suspicious and wishes to stop the call, the system immediately executes this.

In addition, when the user sets the call to be recorded and stored in cloud storage, the system records the call contents and stores them in the user's cloud.

Thereafter, the stored recorded file is utilized to determine whether the voice is a deep voice. For example, the system may analyze this recorded file to recalculate the deep voice probability.

The process of periodically updating the deep voice detection model includes retraining the model using new data reflecting the latest deep voice generation technology. For example, the system collects data on the latest speech synthesis technology and updates the model based thereon to improve performance. When this update process is completed, the system automatically deploys the updated model to the user terminal.

Finally, the system encrypts information related to the deep-voice-suspected voice data, including at least one of a caller number, a call time, a deep voice probability, a user response, and a recorded file, and automatically reports it to the Cyber Safety Bureau of the National Police Agency. For example, the system may make a report including information such as "caller number: 010-1234-5678, call time: 15:30 on October 17, 2024, deep voice probability: 85%."

The user may immediately block a deep-voice-suspected call, or may record the contents of the call to check later. The recorded call contents are securely stored in the user's cloud storage. The recorded file may be utilized to determine whether the voice is a deep voice, and may also be used in the future to improve the deep voice detection model. The system (400) periodically

updates the deep voice detection model. The update includes retraining the model using new data reflecting the latest deep voice generation technology. For example, when a new deep voice generation technology emerges, deep voice data generated using that technology is collected to retrain the model. In addition, the model may be improved by analyzing deep-voice-suspected voice data recorded by users. The updated model is automatically deployed to the user terminal. The system (400) collects information related to the deep-voice-suspected voice data, including at least one of a caller number, a call time, a deep voice probability, a user response, and a recorded file. The collected information is encrypted to protect personal information, and is automatically reported to the Cyber Safety Bureau of the National Police Agency. Through this, the system can contribute to the prevention of and investigation into deep voice crime.

FIG. 7 is a flow diagram illustrating a method for providing an artificial intelligence solution for detecting a deep voice generated by generative artificial intelligence and preventing a related accident according to an embodiment.

Although process steps, method steps, algorithms, and the like are described in a sequential order in the flowchart of FIG. 7, such processes, methods, and algorithms may be configured to operate in any suitable order. In other words, steps of the processes, methods, and algorithms described in various

embodiments of the present invention do not need to be performed in the order described herein. In addition, although some steps are described as being performed asynchronously, in another embodiment, some of these steps may be performed simultaneously. In addition, the illustration of a process by depiction in a drawing does not mean that the illustrated process excludes other variations and modifications thereto, does not mean that any of the illustrated process or steps thereof is essential to one or more of various embodiments of the present invention, and does not mean that the illustrated process is preferred.

In operation S710, a system (e.g., the system (400) of FIG. 4) may collect an analog voice signal under control of a processor (e.g., the processor (420) of FIG. 4). The system (400) may collect an analog voice signal by acquiring a sound source through a plurality of array microphones spatially separated and arranged in a distributed control area. Each array microphone may include a plurality of microphone units arranged to sense the direction of sound.

The processor (420) collects an analog voice signal by controlling a plurality of array microphones spatially separated and arranged in a distributed control area. That is, by recording sound using a plurality of microphones, richer and more accurate voice information is obtained than when using a single microphone.

Each array microphone includes a plurality of microphone units arranged to sense the direction of sound. Through this microphone arrangement, the direction from which sound is coming is identified, and sound in that direction can be collected more clearly. For example, a beamforming technique may be used that identifies the location of a sound source by analyzing time differences among signals collected from a plurality of microphone units. By utilizing this spatial information, the system (400) can effectively remove ambient noise and selectively collect only the desired voice signal.

In operation S720, the system (400) may segment a voice section using a voice activity detection algorithm. The system (400) may convert the collected analog voice signal into a digital voice signal in real time using an analog-to-digital converter, adjust a sampling rate and a bit depth, remove noise from the digital voice signal using at least one of adaptive filtering, spectral subtraction, and beamforming, adjust the volume using automatic gain control, and segment a voice section using a voice activity detection algorithm.

The system segments a voice section using a voice activity detection algorithm. In this step, the system converts the collected analog voice signal into a digital voice signal in real time using an analog-to-digital converter. For example, when the user says "hello," the system converts this voice into a digital signal while adjusting the sampling rate and bit depth to optimize the sound quality.

The higher the sampling rate, the more finely the voice is reproduced, and the greater the bit depth, the more the diversity of tone can be added.

Thereafter, the system removes noise from the digital voice signal using at least one of adaptive filtering, spectral subtraction, and beamforming. For example, in an environment with a lot of background noise, adaptive filtering may be applied to clearly extract only the user's voice.

In addition, the system adjusts the volume using automatic gain control. This is a function that automatically adjusts the volume when the user's voice is too quiet or too loud, in order to maintain a consistent volume. Finally, the system segments a voice section using a voice activity detection algorithm. This algorithm detects the start and end of a voice to clearly distinguish between a section with voice and a section without voice. For example, when the user pauses after saying "hello," the system recognizes this section as a voice section and identifies the subsequent section as a section without voice.

The system (400) converts the collected analog voice signal into a digital voice signal in real time using an analog-to-digital converter (ADC). At this time, the quality of the digital voice signal is determined by adjusting the sampling rate and bit depth. The sampling rate indicates how many times per second the analog signal is measured, and the bit depth indicates the number of bits used to represent the amplitude of each sample. Various techniques are applied to

remove noise from the digital voice signal. Adaptive filtering is a technique that removes noise by automatically adjusting filter coefficients according to the noise characteristics of the surrounding environment. Spectral subtraction is a technique that removes noise by estimating a noise spectrum and subtracting it from the spectrum of the original signal. Beamforming is a technique that uses array microphones to enhance sound in a specific direction and attenuate sound in other directions.

The system (400) adjusts the volume using automatic gain control (AGC).

AGC is a technique that automatically amplifies or attenuates a signal according to the magnitude of the input signal to maintain a constant magnitude of the output signal.

The system (400) segments a voice section using a voice activity detection (VAD) algorithm. The VAD algorithm distinguishes a section in which voice is present from a silent section by analyzing the energy, frequency characteristics, temporal changes, and the like of the voice signal. Through this, the efficiency of speech recognition and analysis can be increased.

In operation S730, the system (400) may calculate a first deep voice probability by weighted-averaging the output results of each model. The system (400) may extract, on a frame-by-frame basis, acoustic features including Mel-Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC)

coefficients, pitch, formant, and a spectrogram from the preprocessed digital voice signal, adjust a frame size and a frame shift interval, and input the extracted acoustic features into a first generative AI model, which is an ensemble AI model including a deep-learning-based acoustic model, a natural-language-processing-based language model, and a statistics-based pattern recognition model, to calculate a first deep voice probability by weighted-averaging the output results of each model.

The system (400) calculates a first deep voice probability by weighted-averaging the output results of each model. In this process, the system extracts several acoustic features from the preprocessed digital voice signal. For example, the system extracts, on a frame-by-frame basis, acoustic features including Mel-Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC) coefficients, pitch, formant, and a spectrogram. Each feature is used to analyze the voice in a different way. For example, MFCC represents the frequency components of the voice, and LPC models the characteristics of the tone.

Pitch represents the pitch of the voice, and formant represents the frequency band of a specific vowel.

A spectrogram visually represents a change in frequency over time, and is useful for analyzing a pronunciation pattern.

The system (400) optimizes each acoustic feature by adjusting the frame size and the frame shift interval. For example, by setting the frame size to 20 ms and the shift interval to 10 ms, detailed analysis can be performed at short time intervals.

Thereafter, the system (400) inputs the extracted acoustic features into a first generative AI model, which is an ensemble AI model including a deep-learning-based acoustic model, a natural-language-processing-based language model, and a statistics-based pattern recognition model. Each model determines whether the voice is a deep voice based on the input acoustic features, and at this time, calculates a first deep voice probability by weighted-averaging the output results of each model. For example, when the acoustic model determines with a probability of 75%, the language model with a probability of 80%, and the pattern recognition model with a probability of 70% that the voice is a deep voice, the system combines these to calculate a final probability.

The system (400) divides the preprocessed digital voice signal into frames, and extracts, from each frame, acoustic features including Mel-Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC) coefficients, pitch, formant, and a spectrogram. The precision of the analysis may be adjusted by adjusting the frame size and the frame shift interval. The extracted acoustic features are input into a first generative AI model, which is an ensemble AI

model. The first generative AI model includes a deep-learning-based acoustic model, a natural-language-processing-based language model, and a statistics-based pattern recognition model. Each model analyzes the input acoustic features to determine whether the voice is a deep voice.

The system (400) calculates a first deep voice probability by weighted-averaging the output results of each model. The weight is determined according to the performance of each model, and a higher weight is given to a model with better performance. Through this, the strengths of each model are combined and their weaknesses are compensated for, thereby increasing the accuracy of deep voice determination.

In operation S740, the system (400) may calculate a second deep voice probability by inputting into a second generative AI model, which is a time-series analysis model. The system (400) may analyze a temporal change including at least one of a pitch change, an energy change, a speech-rate change, a formant change, and a spectral change of the voice signal using the extracted acoustic features, and calculate a second deep voice probability by inputting this into a second generative AI model, which is a time-series analysis model including at least one of a recurrent neural network (RNN)-based model, a long short-term memory (LSTM)-based model, and a gated recurrent unit (GRU)-based model.

The system calculates a final deep voice probability. In this process, the system fuses the first deep voice probability and the second deep voice probability using at least one of weighted averaging, rule-based combination, and machine-learning-based combination. For example, if the first deep voice probability is 70% and the second deep voice probability is 80%, the system may weighted-average these two probabilities to calculate the final probability as 75%. At this time, when the final deep voice probability is equal to or greater than a preset threshold value that is user-configurable and may be dynamically adjusted depending on the situation, the system determines the corresponding voice to be a deep voice and initiates an alert mode. For example, if the user-set threshold value is 75%, the system immediately issues a warning when the final probability exceeds this criterion.

The system (400) analyzes a change pattern over time, including a pitch change, an energy change, a speech-rate change, a formant change, and a spectral change of the voice signal, using the extracted acoustic features. A deep voice, unlike a human voice, may show subtle differences in the temporal change pattern. For example, a deep voice may tend to have a more monotonous pitch change or a more constant speech rate than a human.

The analyzed temporal change information is input into a second generative AI model, which is a time-series analysis model. The second generative AI model

includes at least one of a recurrent neural network (RNN)-based model, a long short-term memory (LSTM)-based model, and a gated recurrent unit (GRU)-based model.

RNN, LSTM, and GRU are all deep learning models specialized for time-series data analysis, and are utilized to analyze current data by remembering information from previous times.

The second generative AI model analyzes the input temporal change information to calculate a second deep voice probability. That is, whether the voice is a deep voice is determined by analyzing how similar the temporal change pattern of the voice signal is to a pattern appearing in a deep voice.

In operation S750, the system (400) may calculate a final deep voice probability. The system (400) may calculate a final deep voice probability by fusing the first deep voice probability and the second deep voice probability using at least one of weighted averaging, rule-based combination, and machine-learning-based combination, and may determine the voice to be a deep voice and initiate an alert mode when the final deep voice probability is equal to or greater than a preset threshold value that is user-configurable and may be dynamically adjusted depending on the situation.

FIG. 8 is a flowchart illustrating a method for detecting a deep voice using signal-quality (SNR)-based adaptive weight allocation and resource optimization according to an embodiment of the present invention.

In step S810, the system (400) extracts, from input voice data, a complex feature vector necessary for deep voice determination. The extracted features include a Mel-Spectrogram representing a characteristic of the frequency domain, MFCC and LPC coefficients representing a temporal change and a characteristic of the vocal tract, and text data including the semantic context of the speech.

In step S820, the system (400) inputs the extracted feature vector into an ensemble model including a plurality of AI models to calculate individual probabilities. According to an embodiment of the present invention, the ensemble model may be composed of a first AI model (acoustic) based on a Convolutional Neural Network (CNN) that analyzes the Mel-Spectrogram, a second AI model (language) based on a Transformer that analyzes the context of the text, and a third AI model (pattern) that analyzes pitch/energy statistics. In particular, in this process, the system (400) performs 'dynamic resource switching' to optimize hardware resources. Specifically, the processor normally activates only the third AI model (pattern recognition), which has a low computational load, to perform a first screening. If a suspicious section in which the deep

voice probability exceeds a preliminary threshold value is not detected as a result of the first analysis, the first AI model (CNN) and the second AI model (Transformer), which occupy a large amount of memory, remain unloaded to minimize power consumption and memory occupancy. If a suspicious section is detected, the processor immediately loads the first and second AI models into active memory to perform a precise analysis.

In step S830, the system (400) measures a signal-to-noise ratio (SNR) by analyzing the signal quality of the input voice data. This is to prevent malfunction of a specific model in an environment with severe ambient noise or poor communication conditions.

In step S840, the system (400) performs 'adaptive weight allocation' that varies, in real time, a weight to be applied to each AI model based on the measured SNR value. For example, when the measured SNR value is a low-quality (noisy) signal that is less than a preset reference value (e.g., 10 dB), the reliability of the first AI model (acoustic), which analyzes fine frequency patterns, may decrease. Accordingly, the system performs a correction that decreases the weight of the first AI model (e.g., w1 = 0.2) and increases the weight of the second AI model (language/context), which is relatively less affected by noise (e.g., w2 = 0.5).

In step S850, the system (400) calculates a final deep voice probability (Pfinal) by computing the determined dynamic weights and the individual probability of each model. The equation applied at this time follows a weighted-average operation of the probability (pi) of each model and the weight (wi) that varies according to the SNR,

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

Through this, precise detection robust to environmental variables is possible.

Finally, when the final deep voice probability exceeds the threshold value (Yes in S860), the system immediately generates an alert or performs a reporting procedure.

Furthermore, as still another embodiment of the present invention, a 'liveness detection feedback' process may be performed after or in parallel with step S860. When a deep voice is suspected, the system (400) presents an additional question (challenge) to the caller through the user terminal, and analyzes the caller's response thereto. At this time, the system measures not only whether the answer is correct, but also the latency from the point in time of the question to the point in time of the response, and the jitter and shimmer variation rate of the response voice waveform, in milliseconds (ms). If the measured latency deviates from the average human response speed (e.g., 200 to 500 ms) (AI

computation delay), or if the jitter variation rate shows an excessively flat pattern (characteristic of an artificial synthesized sound) of less than 0.5%, the system regards this as 'mechanical speech.' Accordingly, even if the content of the answer is accurate, the system adds a penalty weight (e.g., +0.2) to the final deep voice probability to update the probability upward, thereby effectively blocking cleverly trained deep voice attacks.

The final deep voice probability is calculated by fusing the first deep voice probability and the second deep voice probability using at least one of weighted averaging, rule-based combination, and machine-learning-based combination. Weighted averaging is a method of calculating an average by assigning a weight to each probability.

Rule-based combination is a method of combining two probabilities according to predefined rules. For example, a rule such as "if either probability is 0.8 or greater, set the final probability to 0.9" may be used. Machine-learning-based combination is a method of combining two probabilities using a machine learning model. For example, a classification model that takes the two probabilities as input and outputs a final probability may be trained. When the final deep voice probability is equal to or greater than a preset threshold value that is user-configurable and may be dynamically adjusted depending on the situation, the voice is determined to be a deep voice. The threshold value may

be set directly by the user, or may be automatically adjusted by the system depending on the situation. For example, if the user sets a high security level, the threshold value is lowered, increasing the probability of being determined to be a deep voice. Conversely, if the user prioritizes convenience, the threshold value is raised, decreasing the probability of being determined to be a deep voice.

According to an embodiment, when the alert mode is initiated, the system (400) may play a preset warning voice, which may include a message such as "A deep voice has been detected.", "Please be careful.", or "Warning! Stop using deep voice." and is user-configurable, output it to a loudspeaker through digital audio power amplification, display, on a display, a visual warning message that may be displayed in various forms such as text, an image, or an animation simultaneously with the output of the warning voice, generate a vibration or a warning sound in various patterns capable of attracting the user's attention, and transmit, over a network, an alert system including deep-voice-suspected voice data, caller information, a deep voice probability, the current time, location information, a microphone ID, and occurrence time information to a management center host.

When the alert mode is initiated, a warning voice including a message such as "A deep voice has been detected.", "Please be careful.", or "Warning! Stop using

deep voice." is played. A preset voice that is user-configurable may be used as the warning voice. The warning voice is output to a loudspeaker through digital audio power amplification, thereby delivering the warning message to people around the user as well. Simultaneously with the output of the warning voice, a visual warning message that may be displayed in various forms such as text, an image, or an animation is displayed on the display. For example, a text message reading "A deep voice has been detected!" may be displayed on the screen together with a deep voice warning image. The system (400) generates a vibration or a warning sound in various patterns to attract the user's attention. For example, the system may generate a short vibration multiple times, or a warning sound that changes in pitch.

The system (400) transmits, over a network, an alert signal (or alert message) including voice data suspected of being a deep voice, caller information, a deep voice probability, the current time, location information, a microphone ID, and occurrence time information to a management center host. Through this, an administrator can grasp the deep voice occurrence situation in real time and take necessary measures.

According to an embodiment, the system (400) may control to receive alert information from the management center host, display, on an alert display module, caller voice profile information, which may include the caller's gender,

age, region of origin, language used, and voice characteristics, along with the alert occurrence location, and, after receiving the alert information, initiate on-site monitoring to comprehensively analyze sensor data, which may include information collected from CCTV footage, temperature, humidity, illuminance, a motion detection sensor, and the like, and user information, which may include the user's location, authority, access records, and the like, to check on-site information, and, after checking the on-site information, perform voice intercom reconfirmation that supports two-way communication through a microphone and a distributed control area and to which noise removal and voice enhancement technology are applied.

The system (400) receives alert information transmitted from the management center host.

The alert information includes the deep voice occurrence location, the caller's voice profile information, and the like. The caller's voice profile information may include the caller's gender, age, region of origin, language used, voice characteristics, and the like. Such information can help identify and track the deep voice user.

The system (400) displays the received alert information on the alert display module. The alert display module visually displays the deep voice occurrence location, and provides the caller's voice profile information to the user. After

receiving the alert information, the system (400) initiates on-site monitoring. On-site monitoring is a process of comprehensively analyzing sensor data, including information collected from CCTV footage, temperature, humidity, illuminance, a motion detection sensor, and the like, and user information, including the user's location, authority, access records, and the like, to check on-site information.

The system (400) analyzes the collected on-site information to more accurately grasp the deep voice occurrence situation. For example, the location and time at which the deep voice was used can be identified through CCTV footage analysis, and the movement path of the deep voice user can be tracked through motion detection sensor data. After checking the on-site information, the system (400) performs reconfirmation through a voice intercom that supports two-way communication through a microphone and a distributed control area and to which noise removal and voice enhancement technology are applied. That is, the system directly communicates with a person at the site where the deep voice occurred to grasp the situation, and finally verifies whether a deep voice was used.

According to an embodiment, the system controls to receive alert information from the management center host. In this process, the system displays, on the alert display module, caller voice profile information, which may include the caller's

gender, age, region of origin, language used, and voice characteristics. For example, when the user detects a suspicious voice during a call, the system displays information such as "the caller is presumed to be male, appears to be around 30 years old, is from Seoul, and speaks Korean" on the screen to help make a quick decision.

After receiving the alert information, the system initiates on-site monitoring. At this time, the system comprehensively analyzes the situation using sensor data, which may include information collected from CCTV footage, temperature, humidity, illuminance, a motion detection sensor, and the like. For example, movement in a specific space is detected through CCTV, and abnormal signs are detected by grasping environmental conditions through temperature and humidity data. In addition, the system comprehensively analyzes user information, which may include the user's location, authority, access records, and the like, to check on-site information. For example, if a specific user has a history of unauthorized access, the system may issue a warning based on this.

After checking the on-site information, the system performs voice intercom reconfirmation that supports two-way communication through a microphone and a distributed control area and to which noise removal and voice enhancement technology are applied. For example, the voice of the other party can be more clearly delivered by removing background noise that may occur

while the user is on a call. Through this, smooth communication with the user is achieved.

According to an embodiment, the system (400) may, in the voice intercom reconfirmation process, temporarily pause the real-time voice intercom and deliver a warning message reading "Deep voice usage has been detected. Legal liability may follow.", which may be generated using speech synthesis technology and provided in various languages and voices, to the deepfake voice user, may receive information on the deepfake voice user from the management center and issue an alert including the deepfake voice user's location information, risk level, response manual, access route, and surrounding situation information, and related legal information, to notify security personnel, may analyze, based on the deep voice detection result, a user behavior pattern including the frequency of voice usage, time, place, conversation partner, words used, emotional changes, and the like, to predict a potential dangerous situation in advance, and may control to propose warning and preventive measures, such as deep voice prevention education, security system reinforcement, and reinforcement of user authentication procedures, to the user terminal.

In the voice intercom reconfirmation process, the system (400) temporarily pauses the real-time voice intercom and delivers a warning message generated using

speech synthesis technology to the deepfake voice user. The warning message includes content such as "Deep voice usage has been detected. Legal liability may follow.", and may be provided in various languages and voices.

The system (400) receives information on the deepfake voice user from the management center. The deepfake voice user's information includes location information, risk level, response manual, access route, surrounding situation information, and related legal information. The system (400) issues an alert based on the received information to notify security personnel. The system (400) analyzes a user behavior pattern, including the frequency of voice usage, time, place, conversation partner, words used, emotional changes, and the like, based on the deep voice detection result. A potential dangerous situation is predicted in advance through the analyzed behavior pattern. For example, if the frequency of using a specific word at a specific place at a specific time is high, this may be determined to be a potential dangerous situation.

The system (400) proposes warning and preventive measures, such as deep voice prevention education, security system reinforcement, and reinforcement of user authentication procedures, to the user terminal. For example, deep voice prevention educational materials may be provided, or security reinforcement measures such as password changes and two-factor authentication settings may be recommended.

According to an embodiment, the system (400) may collect voice data in real time through a sensor, perform a preprocessing process including noise removal, volume adjustment, and voice-section segmentation on the collected voice data, determine a voice segment and a noise segment in the preprocessed voice data through endpoint detection, remove the noise segment, and perform framing on the voice segment using a Hamming window.

According to an embodiment, the system (400) may perform high-frequency pre-emphasis processing using a first-order high-pass filter for each frame, remove, for the voice data of the voice segment, an effect due to glottal excitation from a fundamental voice signal frequency spectrum, remove an effect of formants through a liftering method to obtain a glottal waveform signal, construct an acoustic model using linear predictive coding and a model of discrete or collected works with precision, and finally obtain the glottal waveform signal using liftering.

According to an embodiment, the system (400) may extract, from the glottal waveform signal, feature parameters including a normalized amplitude, a glottal closure time, a pitch change, a formant change, a spectrogram, and Mel-Frequency Cepstral Coefficients (MFCC), provide the extracted feature parameters as input to an SVM model trained using training data including deep voice data generated by a generative AI model to perform classification,

determine, using the classification result, a speaker's state including one of a normal voice, a deep-voice suspicion, and a deep-voice confirmation, output a speaker state label, and, based on the speaker state label, perform, when a deep-voice-suspicion or deep-voice-confirmation label is received, at least one operation of displaying a warning message, blocking the call, recording, and reporting to the Cyber Safety Bureau of the National Police Agency.

According to an embodiment, the system temporarily pauses the real-time voice intercom in the voice intercom reconfirmation process. At this time, the system delivers a warning message, which may be generated using speech synthesis technology, to the deepfake voice user. For example, the system provides a warning message reading "Deep voice usage has been detected. Legal liability may follow." in various languages and voices. This can be usefully used even in a multinational environment so that the user can understand it. After the warning message is delivered, the system receives information on the deepfake voice user from the management center. This information includes the deepfake voice user's location information, risk level, response manual, access route, surrounding situation information, and related legal information. For example, the system enables security personnel to respond quickly through information such as "risk level: high, access route: entrance 3."

Thereafter, the system issues an alert to security personnel so that they can respond quickly to the situation. The system analyzes a user behavior pattern, including the frequency of voice usage, time, place, conversation partner, words used, emotional changes, and the like, based on the deep voice detection result, to predict a potential dangerous situation in advance. For example, if a pattern is found in which the user frequently makes suspicious statements at a specific time, the system records this and regards it as a risk factor. Finally, the system controls to propose warning and preventive measures, such as deep voice prevention education, security system reinforcement, and reinforcement of user authentication procedures, to the user terminal. For example, the system sends a notification such as "Please participate in an educational program to prevent deep voice usage." to the user to induce preventive measures. Through this, the user's awareness can be raised, contributing to reducing potential risk.

According to an embodiment, the system (400) may emit an acoustic signal in response to a profile processing request associated with a device, receive a customized series of echo acoustic signals based on an acoustic signal reflected from a unique contour of one or more depth portions associated with a user, extract one or more region segments associated with the echo acoustic signals

to train a classification model, and generate a classification model based on the extracted one or more region segments.

In response to a profile processing request associated with a device, the system (400) emits, through a speaker, an acoustic signal having a specific frequency. This acoustic signal is reflected and returned from the unique contour of one or more depth portions of the user's body, particularly the face. The system (400) receives the reflected acoustic signal through a microphone, and converts it into a customized series of echo acoustic signals. The classification model is used to authenticate the user. To train the classification model, the system (400) extracts one or more region segments from the echo acoustic signals. Such region segments include information representing the user's unique physical characteristics. The classification model is trained using the extracted region segments.

For example, a characteristic pattern may be extracted from an echo acoustic signal containing depth information of the user's nose, mouth, ears, and the like, and a classification model that distinguishes between "user A," "user B," and the like may be generated based thereon.

According to an embodiment, the system (400) may extract a combined feature representation based on a classification model, wherein the combined feature representation includes acoustic features and visual landmark features of the

user obtained simultaneously during a discrete epoch, generate a vector-based classification model used to predict the combined feature representation, determine, based on the prediction of the combined feature representation, whether the combined feature representation is associated with an echo signature, wherein the combined feature representation associated with a user profile includes extracted landmark coordinates and acoustic features related to the unique contour of one or more depth portions for the discrete epoch, and, when generating the combined feature representation associated with the user profile, may augment the combined feature representation with synthesized acoustic features and augmented landmark coordinates related to the unique contour of one or more depth portions associated with the user for a changed discrete epoch.

The system (400) extracts a combined feature representation based on the trained classification model. The combined feature representation includes acoustic features and visual landmark features of the user obtained simultaneously during a discrete epoch, that is, a specific period of time. The acoustic features are features extracted from the echo acoustic signal described above. The visual landmark features are location information of the eyes, nose, mouth, and the like extracted from an image of the user's face obtained through a camera.

The system (400) generates a vector-based classification model used to predict the combined feature representation. This classification model determines whether the combined feature representation matches an echo signature, that is, the user's unique acoustic and visual characteristics. The user profile includes extracted landmark coordinates and acoustic features related to the unique contour of one or more depth portions for the discrete epoch. When generating the user profile, the system (400) may augment the combined feature representation using the user's synthesized acoustic features and augmented landmark coordinates for a changed discrete epoch. This is to maintain authentication accuracy in consideration of the fact that the user's appearance or acoustic features may change over time.

According to an embodiment, the system (400) may extract one or more changed vector features associated with synthesized acoustic features and/or augmented landmark coordinates to generate a combined feature representation for a changed discrete epoch. A profile processing request associated with a computing device may include initial registration of an original user profile or user profile authentication for a current epoch for computing device access. A profile processing request associated with a computing device may also include authenticating a current user profile for computing device access by comparing a vector feature associated with a

combined feature representation of the current user profile for the current epoch with a vector feature associated with a combined feature representation of the original user profile.

The system (400) extracts one or more changed vector features associated with synthesized acoustic features and/or augmented landmark coordinates to generate a combined feature representation for a changed discrete epoch. The profile processing request may include initial registration of an original user profile or user profile authentication for a current epoch for computing device access. That is, the system (400) generates a profile when a user is first registered, and thereafter authenticates the user profile for the current epoch each time the user attempts to access the device.

User profile authentication is performed by comparing a vector feature associated with a combined feature representation of the current user profile for the current epoch with a vector feature associated with a combined feature representation of the original user profile. That is, the system (400) compares the current user's features with the registered original user's features to verify whether they match. If the similarity between the two feature vectors is sufficiently high, user authentication succeeds and device access is permitted.

According to an embodiment, the system emits an acoustic signal in response to a profile processing request associated with a device. In this process, the system

generates an acoustic signal of a specific frequency tailored to the user's environment. For example, when the user requests an acoustic signal to listen to music, the system emits a high-quality audio signal to enhance the clarity of the music.

A customized series of echo acoustic signals is received based on an acoustic signal reflected from a unique contour of one or more depth portions associated with the user. For example, when the user emits an acoustic signal in a specific space, an echo acoustic signal, reflected back from a wall or furniture, is collected. This echo signal contains information about the user's location or the structure of the space, and the system analyzes it to enhance its understanding of the environment.

The system extracts one or more region segments associated with the echo acoustic signals to train a classification model. For example, the system identifies a main segment of the signal by analyzing a change in a specific frequency band in the received echo signal. Through this, each segment contains important information about the user's environment or location, and a classification model can be generated based on such segments.

According to an embodiment, the system extracts a combined feature representation based on a classification model. In this process, the system simultaneously analyzes acoustic features and visual landmark features

collected from the user to generate a combined feature representation. For example, when the user speaks in a specific space, image information of that space (e.g., an image captured by a camera) is collected together with the voice.

A vector-based classification model used to predict the combined feature representation is generated. This model is designed to combine and analyze acoustic data and visual data. For example, the system trains the model by combining a frequency component of the voice with landmark information that visually represents the user's location.

Based on the prediction of the combined feature representation, it is determined whether the combined feature representation is associated with an echo signature. For example, the system evaluates the likelihood that a specific acoustic pattern occurs in a specific environment to determine whether this pattern is similar to a previously collected echo signature.

The combined feature representation associated with the user profile includes extracted landmark coordinates and acoustic features related to the unique contour of one or more depth portions for the discrete epoch. For example, the system analyzes a contour based on the user's physical shape or characteristics to generate landmark coordinates for that contour.

When generating the combined feature representation associated with the user profile, the combined feature representation may be augmented with

synthesized acoustic features and augmented landmark coordinates related to the unique contour of one or more depth portions associated with the user for a changed discrete epoch. For example, the system may learn acoustic features generated by the user in a new environment and combine them with existing profile data to generate a more sophisticated user model.

According to an embodiment, the system extracts one or more changed vector features associated with synthesized acoustic features and/or augmented landmark coordinates to generate a combined feature representation for a changed discrete epoch. For example, when the user generates a voice in a new environment, the system collects acoustic features specific to that environment and updates the vector features based thereon.

A profile processing request associated with a computing device may include initial registration of an original user profile or user profile authentication for a current epoch for computing device access. For example, when the user uses a new device, the system checks the original profile information to grant access rights.

A profile processing request associated with a computing device may also include authenticating a current user profile for computing device access by comparing a vector feature associated with a combined feature representation of the current user profile for the current epoch with a vector feature associated with

a combined feature representation of the original user profile. For example, based on an acoustic pattern generated by the user at a specific time and place, the system checks whether this pattern matches the data of the original profile to authenticate the user's identity. Through this process, the system can perform more secure and reliable user authentication.

The embodiments described above may be implemented by a hardware component, a software component, and/or a combination of a hardware component and a software component. For example, the devices, methods, and components described in the embodiments may be implemented using one or more general-purpose or special-purpose computers, such as a processor, a controller, an arithmetic logic unit (ALU), a digital signal processor, a microcomputer, a field programmable gate array (FPGA), a programmable logic unit (PLU), a microprocessor, or any other device capable of executing and responding to instructions.

A processor according to an embodiment of the present invention performs 'dynamic resource switching' for efficient use of computing resources. Normally, only a statistics-based model (the third AI model), which has a low computational load, is activated to perform a first screening, and only when a suspicious section in which the first deep voice probability exceeds a threshold value is detected are the deep learning models (the first and second AI models),

which occupy a large amount of memory, loaded into memory to perform a precise analysis. Through this, hardware optimization is applied that minimizes CPU occupancy and power consumption in an always-on monitoring environment while guaranteeing real-time detection speed (latency).

The processing device may execute an operating system (OS) and one or more software applications executed on the operating system. In addition, the processing device may access, store, manipulate, process, and generate data in response to the execution of software. For convenience of understanding, although it is described in some cases that a single processing device is used, a person having ordinary skill in the art will appreciate that the processing device may include a plurality of processing elements and/or a plurality of types of processing elements. For example, the processing device may include a plurality of processors, or a single processor and a single controller. In addition, other processing configurations, such as a parallel processor, are also possible.

Software may include a computer program, code, an instruction, or a combination of one or more thereof, and may configure a processing device to operate as desired, or may independently or collectively instruct the processing device. Software and/or data may be permanently or temporarily embodied in any type of machine, component, physical device, virtual equipment, computer storage medium or device, or transmitted signal wave, in order to be interpreted by, or

to provide instructions or data to, the processing device. The software may be distributed over networked computer systems, and may be stored or executed in a distributed manner. The software and data may be stored on one or more computer-readable recording media.

Although the embodiments have been described above with reference to the limited drawings, a person having ordinary skill in the art may apply various technical modifications and variations based thereon. For example, appropriate results may be achieved even if the described techniques are performed in an order different from the described method, and/or components of the described system, structure, device, circuit, and the like are combined or combined in a form different from the described method, or replaced or substituted by other components or equivalents.

[Industrial Applicability]

The present invention is applicable to the information and communication security industry, voice phishing prevention systems, and financial security service fields that require artificial-intelligence-based voice analysis technology.

Alternatively, the present invention has industrial applicability in various digital content security industries and communication service fields that require deep voice detection and user identity verification.

WHAT IS CLAIMED IS:

Claim 1.

A system for detecting a deep voice generated by generative artificial intelligence, the system comprising:

a memory configured to store instructions; and one or more processors configured to execute the instructions,

wherein the one or more processors are configured to: extract, from input voice data, a feature vector including a frequency characteristic and a temporal change pattern; input the extracted feature vector into an ensemble artificial intelligence model including a first AI model that is an acoustic model, a second AI model that is a language model, and a third AI model that is a pattern recognition model, to calculate an individual deep voice probability for each model; measure a signal quality of the input voice data to calculate a signal-to-noise ratio (SNR); calculate a final deep voice probability through an adaptive weight allocation algorithm that varies, in real time, a weight to be applied to each AI model based on the calculated SNR value; and generate a deep-voice-suspicion alert when the final deep voice probability exceeds a threshold value.

Claim 2.

The system of claim 1, wherein

the one or more processors are configured to perform dynamic resource switching to control a computing resource load, wherein the input voice data is first analyzed using the

third AI model, of which a computational load is less than a preset reference, and only when a section in which a deep voice probability exceeds a preliminary threshold value is detected as a result of the first analysis are the first AI model and the second AI model, which are deep-learning-based, loaded into an active area of the memory to perform a precise analysis.

Claim 3.

The system of claim 1, wherein

the one or more processors are configured to: when the final deep voice probability exceeds the threshold value, present, through a user terminal, an additional question for verifying a caller's identity; receive a caller's response voice signal to the additional question, and measure a latency from a point in time of the question to a point in time of the response and a jitter and shimmer variation rate of the response voice waveform; and, when the latency deviates from a preset human response range or the jitter variation rate falls within a preset artificial-synthesized-sound pattern range, perform liveness detection feedback that adds a penalty weight to the final deep voice probability and updates the final deep voice probability, regardless of the content of the response.

Claim 4.

The system of claim 1, wherein

the adaptive weight allocation algorithm is configured to, when the SNR value is a low-quality signal that is less than a reference value, decrease a weight of the first AI model (acoustic), which is sensitive to noise, and increase a weight of the second AI model (language),

which is context-dependent, and calculate the final deep voice probability (Pfinal) based on the following equation, where wi is a weight of each model that varies according to the SNR, and pi is an output probability of each model,

_

Claim 5.

The system of claim 1, wherein

the first AI model is a Convolutional Neural Network (CNN)-based model that receives a Mel-Spectrogram as input, the second AI model is a Transformer-based language model that receives speech content converted into text as input, and the third AI model is a pattern recognition model that receives pitch and energy change statistics of the voice as input.

Claim 6.

A method for detecting a deep voice, the method being performed by a computing device, the method comprising:

extracting, from input voice data, a feature vector including a frequency characteristic and a temporal change pattern;

inputting the extracted feature vector into an ensemble model including a first AI model that is an acoustic model, a second AI model that is a language model, and a third AI model that is a pattern recognition model, to calculate an individual probability;

measuring a signal-to-noise ratio (SNR) by analyzing a signal quality of the input voice data;

calculating a final deep voice probability by performing adaptive weight allocation that dynamically determines a weight to be applied to each of the plurality of AI models according to the measured SNR value; and

when the final deep voice probability exceeds a threshold value, displaying a warning message on a user terminal or transmitting a report signal to an investigative agency server.

Claim 7.

A computer-readable recording medium having recorded thereon a program for executing the method of claim 6 on a computer.