(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) **World Intellectual Property Organization**
International Bureau

(43) **International Publication Date**
03 January 2019 (03.01.2019)

WIPO | PCT

(10) **International Publication Number**
WO 2019/002849 A1

(51) **International Patent Classification:**
G10L 17/02 (2013.01) ***H04L 9/32*** (2006.01)
G10L 25/03 (2013.01) ***H04R 29/00*** (2006.01)
G06F 21/32 (2013.01) ***H04W 12/00*** (2009.01)

(21) **International Application Number:**
PCT/GB2018/051787

(22) **International Filing Date:**
27 June 2018 (27.06.2018)

(25) **Filing Language:** English

(26) **Publication Language:** English

(30) **Priority Data:**
62/526,013 28 June 2017 (28.06.2017) US

(71) **Applicant: CIRRUS LOGIC INTERNATIONAL SEMICONDUCTOR LIMITED** [GB/GB]; 7B Nightingale Way, Quartermile, Edinburgh, Central Scotland EH3 9EG (GB).

(72) **Inventor: LESSO, John Paul**; 8 Galachlaw Shot, Edinburgh, Lothian EH10 7JF (GB).

(74) **Agent: HASELTINE LAKE LLP**; Redcliff Quay, 120 Redcliff Street, Bristol BS1 6HU (GB).

(81) **Designated States** *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

(54) **Title:** DETECTION OF REPLAY ATTACK

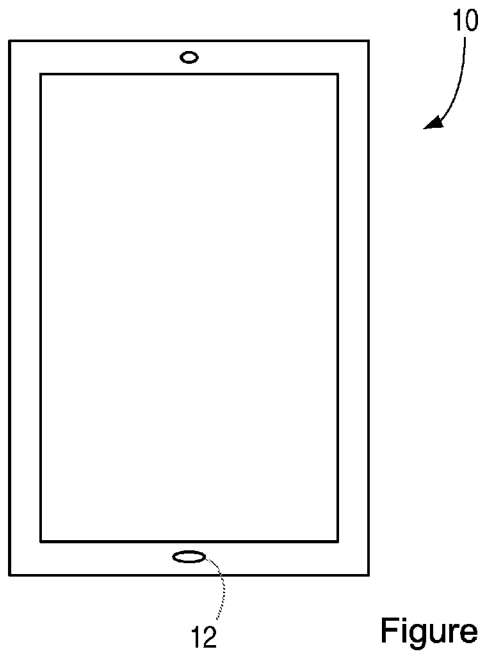

Figure 1

(57) **Abstract:** In order to detect a replay attack in a speaker recognition system, at least one feature is identified in a detected magnetic field. It is then determined whether the at least one identified feature of the detected magnetic field is indicative of playback of speech through a loudspeaker. If so, it is determined that a replay attack may have taken place.

WO 2019/002849 A1

[Continued on next page]

Published:
— *with international search report (Art. 21(3))*

DETECTION OF REPLAY ATTACK

Technical Field

Embodiments described herein relate to methods and devices for detecting a replay attack on a voice biometrics system.

Background

Voice biometrics systems are becoming widely used. In such a system, a user trains the system by providing samples of their speech during an enrolment phase. In subsequent use, the system is able to discriminate between the enrolled user and non-registered speakers. Voice biometrics systems can in principle be used to control access to a wide range of services and systems.

One way for a malicious party to attempt to defeat a voice biometrics system is to obtain a recording of the enrolled user's speech, and to play back the recording in an attempt to impersonate the enrolled user and to gain access to services that are intended to be restricted to the enrolled user.

This is referred to as a replay attack, or as a spoofing attack.

Summary

According to an aspect of the present invention, there is provided a method of detecting a replay attack in a speaker recognition system. The method comprises: identifying at least one feature of a detected magnetic field; determining whether the at least one identified feature of the detected magnetic field is indicative of playback of speech through a loudspeaker; and if so, determining that a replay attack may have taken place.

The method may further comprise: receiving an audio signal representing speech, wherein the audio signal is received at substantially the same time as the magnetic field is detected; and if it is determined that the at least one identified feature of the detected magnetic field is indicative of playback of speech through a loudspeaker, determining that the audio signal may result from said replay attack.

The step of identifying at least one feature of the detected magnetic field may comprise: receiving a signal from a magnetometer; and performing a Discrete Fourier Transform on the received signal from the magnetometer.

The step of identifying at least one feature of the detected magnetic field may comprise: receiving a signal from a magnetometer; and detecting modulation of the at least one feature from the received signal from the magnetometer at frequencies in the range of 2Hz-10Hz, at the syllabic rate, and/or at the articulation rate. The syllabic rate and/or the articulation rate may correspond to typical rates for speech or for that specific type of speech or speaker, or may be determined by analysis of speech that is detected at the same time as the magnetic field is being detected.

The method may comprise determining whether a detected pattern of variability of the magnetic field is indicative of a reference pattern associated with playback of a predetermined spoken phrase through a loudspeaker.

Determining whether the detected pattern of variability of the magnetic field is indicative of a reference pattern associated with playback of a predetermined spoken phrase through a loudspeaker may comprise comparing the detected pattern of variability of the magnetic field with a stored reference pattern.

The stored reference pattern may correspond to the predetermined spoken phrase, as spoken by a specific enrolled user, or may correspond to the predetermined spoken phrase, as spoken by multiple speakers.

Determining whether the detected pattern of variability of the magnetic field is indicative of a reference pattern associated with playback of a predetermined spoken phrase through a loudspeaker may comprise passing the detected pattern of variability of the magnetic field to a classifier that has been trained to with inputs obtained from playback of the predetermined spoken phrase through a loudspeaker.

According to a second aspect of the present invention, there is provided a system for detecting a replay attack in a speaker recognition system, the system being configured for: identifying at least one feature of a detected magnetic field; determining whether the at least one identified feature of the detected magnetic field is indicative of playback of speech through a loudspeaker; and if so, determining that a replay attack may have taken place.

The system may be further configured for: receiving an audio signal representing speech, wherein the audio signal is received at substantially the same time as the magnetic field is detected; and if it is determined that the at least one identified feature of the detected magnetic field is indicative of playback of speech through a loudspeaker, determining that the audio signal may result from said replay attack.

Identifying at least one feature of the detected magnetic field may comprise: receiving a signal from a magnetometer; and performing a Discrete Fourier Transform on the received signal from the magnetometer.

The system may be configured for identifying at least one feature of the detected magnetic field by: receiving a signal from a magnetometer; and detecting modulation of the at least one feature from the received signal from the magnetometer at frequencies in the range of 2Hz-10Hz.

The system may be configured for identifying at least one feature of the detected magnetic field by: receiving a signal from a magnetometer; and detecting modulation of the at least one feature from the received signal from the magnetometer at the syllabic rate.

The system may be configured for identifying at least one feature of the detected magnetic field by: receiving a signal from a magnetometer; and detecting modulation of the at least one feature from the received signal from the magnetometer at the articulation rate.

The system may be configured for determining whether a detected pattern of variability of the magnetic field is indicative of a reference pattern associated with playback of a predetermined spoken phrase through a loudspeaker.

The system may be configured for determining whether the detected pattern of variability of the magnetic field is indicative of a reference pattern associated with playback of a predetermined spoken phrase through a loudspeaker by: comparing the detected pattern of variability of the magnetic field with a stored reference pattern.

The stored reference pattern may correspond to the predetermined spoken phrase, as spoken by a specific enrolled user, or may correspond to the predetermined spoken phrase, as spoken by multiple speakers.

The system may be configured for determining whether the detected pattern of variability of the magnetic field is indicative of a reference pattern associated with playback of a predetermined spoken phrase through a loudspeaker by: passing the detected pattern of variability of the magnetic field to a classifier that has been trained to with inputs obtained from playback of the predetermined spoken phrase through a loudspeaker.

According to an aspect of the present invention, there is provided a device comprising a system according to the second aspect. The device may comprise a mobile telephone, an audio player, a video player, a mobile computing platform, a games device, a remote controller device, a toy, a machine, or a home automation controller or a domestic appliance.

According to an aspect of the present invention, there is provided a computer program product, comprising a computer-readable tangible medium, and instructions for performing a method according to the first aspect.

According to an aspect of the present invention, there is provided a non-transitory computer readable storage medium having computer-executable instructions stored thereon that, when executed by processor circuitry, cause the processor circuitry to perform a method according to the first aspect.

According to an aspect of the present invention, there is provided a device comprising the non-transitory computer readable storage medium according to the previous aspect. The device may comprise a mobile telephone, an audio player, a video player, a mobile computing platform, a games device, a remote controller device, a toy, a machine, or a home automation controller or a domestic appliance.

Brief Description of Drawings

For a better understanding of the present invention, and to show how it may be put into effect, reference will now be made to the accompanying drawings, in which:-

Figure 1 illustrates a smartphone.

Figure 2 is a schematic diagram, illustrating the form of the smartphone.

Figure 3 illustrates a first situation in which a replay attack is being performed;

Figure 4 illustrates a second situation in which a replay attack is being performed;

Figure 5 is a block diagram of a speech processing system;

Figure 6 is a flow chart illustrating a method in accordance with the invention;

Figure 7 is a block diagram of a system for implementing one method;

Figure 8 is a block diagram illustrating a first element of the system of Figure 7;

Figure 9 is a block diagram illustrating an alternative form of the first element of the system of Figure 7;

Figure 10 is a block diagram of a system for implementing a part of the method;

Figure 11 is a block diagram of a system for implementing a part of the method;

Figure 12 is a block diagram of a system for implementing a part of a method;

Figure 13 is a block diagram illustrating a speech processing system;

Figure 14 is a flow chart illustrating a method performed in the system of Figure 13; and

Figure 15 shows a possible form of a block in the system of Figure 13.

Detailed Description of Embodiments

The description below sets forth example embodiments according to this disclosure. Further example embodiments and implementations will be apparent to those having ordinary skill in the art. Further, those having ordinary skill in the art will recognize that various equivalent techniques may be applied in lieu of, or in conjunction with, the embodiments discussed below, and all such equivalents should be deemed as being encompassed by the present disclosure.

Figure 1 illustrates a smartphone 10, having a microphone 12 for detecting ambient sounds. In normal use, the microphone is of course used for detecting the speech of a user who is holding the smartphone 10.

Figure 2 is a schematic diagram, illustrating the form of the smartphone 10.

Specifically, Figure 2 shows various interconnected components of the smartphone 10. It will be appreciated that the smartphone 10 will in practice contain many other components, but the following description is sufficient for an understanding of the present invention.

Thus, Figure 2 shows the microphone 12 mentioned above. In certain embodiments, the smartphone 10 is provided with multiple microphones 12, 12a, 12b, etc.

Figure 2 also shows a memory 14, which may in practice be provided as a single component or as multiple components. The memory 14 is provided for storing data and program instructions.

Figure 2 also shows a processor 16, which again may in practice be provided as a single component or as multiple components. For example, one component of the processor 16 may be an applications processor of the smartphone 10.

Figure 2 also shows a transceiver 18, which is provided for allowing the smartphone 10 to communicate with external networks. For example, the transceiver 18 may include circuitry for establishing an internet connection either over a WiFi local area network or over a cellular network.

Figure 2 also shows audio processing circuitry 20, for performing operations on the audio signals detected by the microphone 12 as required. For example, the audio processing circuitry 20 may filter the audio signals or perform other signal processing operations.

Figure 2 also shows at least one sensor 22. In embodiments of the present invention, the sensor is a magnetic field sensor for detecting a magnetic field. For example, the sensor 22 may be a Hall effect sensor, that is able to provide separate measurements of the magnet field strength in three orthogonal directions.

In this embodiment, the smartphone 10 is provided with voice biometric functionality, and with control functionality. Thus, the smartphone 10 is able to perform various functions in response to spoken commands from an enrolled user. The biometric functionality is able to distinguish between spoken commands from the enrolled user, and the same commands when spoken by a different person. Thus, certain embodiments of the invention relate to operation of a smartphone or another portable electronic device with some sort of voice operability, for example a tablet or laptop computer, a games console, a home control system, a home entertainment system, an in-vehicle entertainment system, a domestic appliance, or the like, in which the voice biometric functionality is performed in the device that is intended to carry out the spoken command. Certain other embodiments relate to systems in which the voice biometric functionality is performed on a smartphone or other device, which then transmits the commands to a separate device if the voice biometric functionality is able to confirm that the speaker was the enrolled user.

In some embodiments, while voice biometric functionality is performed on the smartphone 10 or other device that is located close to the user, the spoken commands are transmitted using the transceiver 18 to a remote speech recognition system, which determines the meaning of the spoken commands. For example, the speech recognition system may be located on one or more remote server in a cloud computing environment. Signals based on the meaning of the spoken commands are then returned to the smartphone 10 or other local device.

One attempt to deceive a voice biometric system is to play a recording of an enrolled user’s voice in a so-called replay or spoof attack.

Figure 3 shows an example of a situation in which a replay attack is being performed. Thus, in Figure 3, the smartphone 10 is provided with voice biometric functionality. In this example, the smartphone 10 is in the possession, at least temporarily, of an attacker, who has another smartphone 30. The smartphone 30 has been used to record the voice of the enrolled user of the smartphone 10. The smartphone 30 is brought close to the microphone inlet 12 of the smartphone 10, and the recording of the enrolled user's voice is played back. If the voice biometric system is unable to detect that the enrolled user's voice that it detects is a recording, the attacker will gain access to one or more services that are intended to be accessible only by the enrolled user.

It is known that smartphones, such as the smartphone 30, are typically provided with loudspeakers that are of relatively low quality due to size constraints. Thus, the recording of an enrolled user's voice played back through such a loudspeaker will not be a perfect match with the user's voice, and this fact can be used to identify replay attacks. For example, loudspeakers may have certain frequency characteristics, and if these frequency characteristics can be detected in a speech signal that is received by the voice biometrics system, it may be considered that the speech signal has resulted from a replay attack.

Figure 4 shows a second example of a situation in which a replay attack is being performed, in an attempt to overcome the method of detection described above. Thus, in Figure 4, the smartphone 10 is provided with voice biometric functionality. Again, in this example, the smartphone 10 is in the possession, at least temporarily, of an attacker, who has another smartphone 40. The smartphone 40 has been used to record the voice of the enrolled user of the smartphone 10.

In this example, the smartphone 40 is connected to a high quality loudspeaker 50. Then, the microphone inlet 12 of the smartphone 10 is positioned close to the loudspeaker 50, and the recording of the enrolled user's voice is played back through the loudspeaker 50. As before, if the voice biometric system is unable to detect that the enrolled user's voice that it detects is a recording, the attacker will gain access to one or more services that are intended to be accessible only by the enrolled user.

In this example, the loudspeaker 50 may be of high enough quality that the recording of the enrolled user's voice played back through the loudspeaker will not be reliably distinguishable from the user's voice, and so the audio features of the speech signal cannot be used to identify the replay attack.

However, it is appreciated that many loudspeakers, and particularly high quality loudspeakers, are electromagnetic loudspeakers in which an electrical audio signal is applied to a voice coil, which is located between the poles of a permanent magnet, causing the coil to move rapidly backwards and forwards. This movement causes a diaphragm attached to the coil to move backwards and forwards, creating sound waves. It is recognised here that, if a device such as the smartphone 10 is positioned close to a loudspeaker while it is playing back sounds, there will be corresponding changes in the magnetic field, which will be detectable by a magnetic field sensor 22.

Figure 5 illustrates a part of a speech processing system 60, for use in a device such as a smartphone 10 as shown in Figure 2. Specifically, Figure 5 shows an input 62, for receiving an audio signal, for example from one or more microphone 12 in a smartphone 10 as shown in Figure 2.

In order to reduce power consumption, the speech processing system 60 operates in a progressive fashion. Thus, the audio signal is passed to a voice activity detection (VAD) block 64, which determines when the received audio signal contains a human voice.

When the VAD block 64 determines that the received audio signal contains human speech, it sends a signal to a keyword detection block 66 to initiate operation thereof.

When it is activated, the keyword detection block 66 receives the input audio signal, and determines whether the human speech contains a predetermined trigger phrase. For example, a smartphone might have a trigger phrase "hello phone", which the user must speak to activate the speech processing.

When the keyword detection block 66 determines that the human speech contains a predetermined trigger phrase, it sends a signal to a speaker recognition block 68 to initiate operation thereof.

When it is activated, the speaker recognition block 68 receives the input audio signal, and determines whether the human speech was spoken by an enrolled user of the device. For example, a smartphone may have just one or a few enrolled users, who are authorised to issue voice commands to the device, and the speaker recognition block 68 determines whether the detected human speech was spoken by that enrolled user or one of the enrolled users.

When the speaker recognition block 68 determines that the human speech was spoken by an enrolled user of the device it sends a signal to a speech processing block 70 to initiate operation thereof. The speech processing block 70 may be located in the same device as the other blocks shown in Figure 5, or it may be located remotely in the cloud.

When it is activated, the speech processing block 70 receives the input audio signal, and determines the content of the received speech. For example, the speech processing block 70 may determine that the speech contains a command, and may then control some aspect of the operation of the device, or of a separate device, in response to that command.

As discussed above with reference to Figure 4, one possible attack on such a system is that an attacker may play back a recording of the enrolled user speaking the predetermined trigger phrase. Without any system for detecting such a replay attack, the attacker may be able to issue commands that would be acted upon by the speech processing system 60.

Figure 6 is a flow chart, illustrating a method of detecting a replay attack on a voice biometrics system, and **Figure 7** is a block diagram illustrating functional blocks in the voice biometrics system. Specifically, Figure 7 illustrates functional blocks in the keyword detection block 66.

Specifically, in step 80 in the method of Figure 6, an audio signal is received on an input 100 of the keyword detection block 66 shown in Figure 7. For example, in a device as shown in Figure 2, the audio signal received on the input 100 may be the audio signal detected by the microphone 12, or may be the sum of the audio signals detected by the microphones if there is more than one.

At the same time, in step 82 in the method of Figure 6, an input signal is received on an input 102 of the system shown in Figure 7. The input signal received on the input 102 is received from a magnetometer. For example, when the method is performed in a device such as a smartphone or a tablet computer, the device will typically include a three-axis magnetometer, which generates an output signal containing separate measurements of the magnetic field strength in three orthogonal directions.

In some embodiments, the input signal received from the magnetometer is passed to a magnetic feature extraction block 104. For example, if the signal received from the magnetometer contains separate measurements of the magnetic field strength in three orthogonal directions, these can be combined to provide a single measurement of the magnetic field strength. The measurement of the magnetic field strength could be found as the square root of the sum of the squares of the three separate measurements of the magnetic field strength in the three orthogonal directions.

Further, the aim of the system is to determine any magnetic field that is generated by a nearby object such as a loudspeaker. In order to obtain the most useful information about this, one possibility is to process the input signal received from the magnetometer in order to remove the effects of the Earth's magnetic field. For example, this can be achieved by forming an average value of the magnetic field strength, for example over a period of seconds, minutes or hours, and subtracting this from each individual measurement to obtain an instantaneous measurement of the magnetic field generated by artificial sources.

Thus, in step 82 in the method of Figure 6, after subtracting the static magnetic field from the magnetic field detected by the magnetometer, what is left is a pattern of variability of the magnetic field. If this occurs at substantially the same time as an audio signal is being received on the input 100 of the keyword detection block 66, then the pattern of variability of the magnetic field is considered to be associated in some way with the audio signal.

Relevant information about the pattern of variability of the magnetic field can then be obtained by the magnetic feature extraction block 104.

Figure 8 is a block diagram illustrating a first form of the magnetic feature extraction block 104. Specifically, Figure 8 shows the magnetometer signal, possibly after pre-processing to remove the effects of the Earth's magnetic field and obtain an instantaneous measurement of the magnetic field generated by artificial sources, being applied to a Discrete Fourier Transform (DFT) block 150. The output of the DFT block 150 may then optionally be normalised in a normalisation block 152. The normalisation consists of modifying at least one of the moments of the signal. For example, the first moment (mean) may be set to zero and the second moment (i.e. the standard deviation) may be set to unity.

Figure 9 is a block diagram illustrating a second form of the magnetic feature extraction block 104, in which a robust hash is performed. A robust hash is essentially a lossy compression function that gives similar scores for similar, but not necessarily identical, files. Specifically, Figure 9 shows the magnetometer signal, possibly after pre-processing to remove the effects of the Earth's magnetic field and obtain an instantaneous measurement of the magnetic field generated by artificial sources, being applied to a Discrete Fourier Transform (DFT) block 160, which is used to filter the magnetometer signal. The output of the DFT block 160 is passed to statistics estimation block 162, which is used to estimate the statistics per band, and these are passed to a quantization block 164.

In these examples, a Discrete Fourier Transform is used, though other techniques such as Fast Fourier Transform (FFT) or Discrete Cosine Transform (DCT) can also be used.

If a pattern of variability of the magnetic field that is associated with an audio signal is detected, the process passes to step 84 of the method of Figure 6, in which it is determined whether the detected pattern of variability of the magnetic field is indicative of a reference pattern associated with playback of a predetermined spoken phrase through a loudspeaker. Thus, in the system of Figure 7, the detected pattern of variability of the magnetic field is passed to a determination block 106.

The determination may take any suitable form.

Figure 10 is a block diagram illustrating a first possible form of the determination block 106. Specifically, Figure 8 shows the detected pattern of variability of the magnetic field, that is, the features extracted from the magnetometer signal by the feature extraction block 104, being passed to a first input of a comparison block 120 in the determination block 106. A store 122 is connected to a second input of the comparison block 120.

The store 122 may store a reference pattern, and this reference pattern may correspond to the pattern of variability of the magnetic field when the predetermined spoken phrase is spoken by a specific enrolled user, or may correspond to the average pattern of variability of the magnetic field when the predetermined spoken phrase is spoken by multiple speakers.

Figure 11 is a block diagram illustrating a second possible form of the determination block 106. Specifically, Figure 11 shows the detected pattern of variability of the magnetic field being passed to a classifier block 130 in the determination block 106. The classifier block 130 operates to determine whether the detected pattern of variability of the magnetic field is indicative of a reference pattern of variability of a magnetic field that is associated with playback of the predetermined spoken phrase through a loudspeaker.

In these examples, features are extracted from the magnetometer signal, and compared with a reference pattern, either directly or in a classifier. In other examples, the magnetometer signal itself is passed to a suitably trained classifier to determine whether the signal contains features that indicate that the pattern of variability of the magnetic field is indicative of a reference pattern associated with playback of the predetermined spoken phrase through a loudspeaker.

To avoid the need to train the classifier 130 by playing back many examples of speech through a loudspeaker, the classifier may be trained using modelled data.

Figure 12 is a block diagram showing the training process.

The signal that is received from the magnetometer signal will typically have a sample rate in the region of 80-120Hz. Therefore, if that signal does result from the playback of speech, it will represent a heavily under-sampled version of that speech. Further, the sample rate may not be constant, because the operation of the magnetometer is typically a low priority task in a device such as a smartphone.

Therefore, to obtain suitable training data for the classifier, signals are obtained that represent the voltage applied to a loudspeaker when the predetermined phrase is spoken.

These signals are input, in turn, to a transducer model 140, which maps the voltage applied to the loudspeaker to the magnetic field that the magnetometer senses. The transducer model must therefore take account of different types of loudspeaker that might plausibly be used in a spoof attack, and must also take account of different possible positions of the magnetometer relative to the loudspeaker.

The resulting signals are applied to a magnetometer model 142, which models the sampling process of the magnetometer (namely, the fact that the magnetometer may perform under-sampling, irregular sampling, or other possible sampling techniques).

This provides data, representing many possible magnetometer signals that can be obtained by playing back the predetermined phrase through a loudspeaker and detecting the resulting pattern of variability of the magnetic field through a magnetometer. This obtained data can then be used to train the classifier 130.

As shown in Figure 11, the trained classifier can then be used to determine whether a newly detected pattern of variability of the magnetic field is likely to have been the result of the predetermined phrase being played back through a loudspeaker.

If it is determined in step 84 of the method of Figure 6 that the detected pattern of variability of the magnetic field is indicative of a reference pattern associated with playback of a predetermined spoken phrase through a loudspeaker, the process passes to step 86, in which it is determined that the received audio signal may be associated with a replay attack.

This determination may be used on its own to determine that the received audio signal results from a replay attack. Alternatively, the determination may be combined with other factors to reach a decision as to whether the received audio signal results from a replay attack.

In the system of Figure 7, the audio signal received on the input 100 of the keyword detection block 66 is passed to an audio processing block 108.

In most cases, the audio processing block 108 of the keyword detection block 66 determines whether the audio signal contains a predetermined trigger phrase and, if so, it sends a signal to the speaker recognition block 68 of Figure 5 to initiate operation thereof.

However, if the determination block 106 determines (based only on the pattern of variability of the magnetic field or based partly on the pattern of variability of the magnetic field) that the received audio signal may result from a replay attack shown in Figure 7, then the determination block 106 may send a signal to the audio processing block 108 in order to prevent its operation.

Thus, if it is determined from the magnetic field measurements that the audio signal may result from a replay attack, the keyword detection block 66 does not attempt to detect the presence of the predetermined trigger phrase.

The method has been described so far herein with reference to a specific example in which it is determined whether the detected pattern of variability of the magnetic field is indicative of a reference pattern associated with playback of a predetermined spoken phrase through a loudspeaker, and there is only one such predetermined phrase, for example a trigger phrase that is used by an enrolled user to activate a device. However, there may be multiple predetermined spoken phrases, and the method can test whether the detected pattern of variability of the magnetic field is indicative of a reference pattern associated with playback of any of these predetermined spoken phrases through a loudspeaker. Where the determination block 106 includes a classifier, as shown in Figure 11, this requires more training data for the classifier.

More generally, the method can test whether the detected pattern of variability of the magnetic field is indicative of playback of human speech.

Figure 13 is a block diagram illustrating a speech processing system 170, and **Figure 14** is a flow chart illustrating a method performed in the system.

An audio signal that may contain speech is passed to a speech processing block 172. This may take any suitable form. For example, it may be a keyword detection block, a speaker recognition function, a speech recognition block, or any other function.

A magnetometer signal, possibly after pre-processing to remove the effects of the Earth's magnetic field and obtain an instantaneous measurement of the magnetic field generated by artificial sources, is applied to a magnetic feature extraction block 174.

Thus, in step 180 of the process shown in Figure 180, at least one feature of the detected magnetic field is identified.

The magnetic feature extraction block 174 may for example extract Mel Frequency Cepstral Coefficients (MFCCs) from the magnetometer signal.

Alternatively, **Figure 15** shows a possible form of the magnetic feature extraction block 174. This operates by looking for features that are characteristic of speech, which is typically modulated at the syllabic rate or articulation rate, which may for example be in the region of 1-15Hz, and more specifically in the region of 2-10Hz, and is typically around 4Hz.

In Figure 15, the magnetometer signal is filtered into multiple bands. Figure 15 shows this filtering being performed by a Discrete Fourier Transform (DFT) block 190, but alternatively a filter bank may be used.

The filtered signal is passed to a block 192 in which the signal is converted to energy, by squaring it.

Each energy band is then passed to a band pass filter (BPF) block 194, with a pass band centred on a suitable syllabic rate, for example 4Hz. The syllabic rate, or articulation rate, may be chosen to correspond to a typical rate, for example for general speech or for that specific type of speech or speaker. Alternatively, a value of a syllabic rate, or articulation rate, may be determined by analysis of speech that is detected at the same time as the magnetic field is being detected (for example in the speech processing block 172 of Figure 13). That rate may then be used to set the pass band of the band pass filter (BPF) block 194

The modulation energy at the output of the BPF can then be measured, for example with a simple threshold or with a more advanced pattern recogniser such as neural net.

In block 176 of the system shown in Figure 13, and in step 182 of the method shown in Figure 14, it is then determined whether the at least one identified feature of the detected magnetic field is indicative of playback of speech through a loudspeaker. This may involve determining whether the degree of modulation of the magnetometer signal at the syllabic rate is indicative of playback of speech through a loudspeaker.

If so, then it is determined in step 184 that a replay attack may have taken place. In that event, any suitable output may be provided to a user.

In the case of a system as shown in Figure 13, the output may be used to gate the speech processing, such that the received audio signal is not processed as planned, because it is assumed that it may be the result of a replay attack.

There are therefore disclosed methods and systems that can be used for detecting situations that may indicate that a received audio signal is the result of a replay attack.

The skilled person will recognise that some aspects of the above-described apparatus and methods may be embodied as processor control code, for example on a non-volatile carrier medium such as a disk, CD- or DVD-ROM, programmed memory such as read only memory (Firmware), or on a data carrier such as an optical or electrical signal carrier. For many applications embodiments of the invention will be implemented on a DSP (Digital Signal Processor), ASIC (Application Specific Integrated Circuit) or FPGA (Field Programmable Gate Array). Thus the code may comprise conventional program code or microcode or, for example code for setting up or controlling an ASIC or FPGA. The code may also comprise code for dynamically configuring re-configurable apparatus such as re-programmable logic gate arrays. Similarly the code may comprise code for a hardware description language such as Verilog TM or VHDL (Very high speed integrated circuit Hardware Description Language). As the skilled person will appreciate, the code may be distributed between a plurality of coupled components in communication with one another. Where appropriate, the embodiments may also be implemented using code running on a field-(re)programmable analogue array or similar device in order to configure analogue hardware.

Note that as used herein the term module shall be used to refer to a functional unit or block which may be implemented at least partly by dedicated hardware components such as custom defined circuitry and/or at least partly be implemented by one or more software processors or appropriate code running on a suitable general purpose processor or the like. A module may itself comprise other modules or functional units. A module may be provided by multiple components or sub-modules which need not be co-located and could be provided on different integrated circuits and/or running on different processors.

Embodiments may be implemented in a host device, especially a portable and/or battery powered host device such as a mobile computing device for example a laptop or tablet computer, a games console, a remote control device, a home automation controller or a domestic appliance including a domestic temperature or lighting control system, a toy, a machine such as a robot, an audio player, a video player, or a mobile telephone for example a smartphone.

It should be noted that the above-mentioned embodiments illustrate rather than limit the invention, and that those skilled in the art will be able to design many alternative embodiments without departing from the scope of the appended claims. The word “comprising” does not exclude the presence of elements or steps other than those listed in a claim, “a” or “an” does not exclude a plurality, and a single feature or other unit may fulfil the functions of several units recited in the claims. Any reference numerals or labels in the claims shall not be construed so as to limit their scope.

CLAIMS

1. A method of detecting a replay attack in a speaker recognition system, the method comprising:

identifying at least one feature of a detected magnetic field;

determining whether the at least one identified feature of the detected magnetic field is indicative of playback of speech through a loudspeaker; and

if so, determining that a replay attack may have taken place,

wherein the step of identifying at least one feature of the detected magnetic field comprises:

receiving a signal from a magnetometer; and

detecting modulation of the at least one feature from the received signal from the magnetometer at one or more of:

at least one frequency in the range of 2Hz-10Hz;

a speech syllabic rate; and/or

an articulation rate.

2. A method as claimed in claim 1, further comprising:

receiving an audio signal representing speech, wherein the audio signal is received at substantially the same time as the magnetic field is detected; and

if it is determined that the at least one identified feature of the detected magnetic field is indicative of playback of speech through a loudspeaker, determining that the audio signal may result from said replay attack.

3. A method as claimed in claim 1 or 2, wherein the step of identifying at least one feature of the detected magnetic field comprises:

receiving a signal from a magnetometer; and

performing a Discrete Fourier Transform on the received signal from the magnetometer.

.4. A method as claimed in claim 1, 2 or 3, comprising determining whether a detected pattern of variability of the magnetic field is indicative of a reference pattern associated with playback of a predetermined spoken phrase through a loudspeaker.

5. A method as claimed in claim 4, wherein determining whether the detected pattern of variability of the magnetic field is indicative of a reference pattern associated with playback of a predetermined spoken phrase through a loudspeaker comprises:

comparing the detected pattern of variability of the magnetic field with a stored reference pattern.

6. A method as claimed in claim 5, wherein the stored reference pattern corresponds to the predetermined spoken phrase, as spoken by a specific enrolled user.

7. A method as claimed in claim 5, wherein the stored reference pattern corresponds to the predetermined spoken phrase, as spoken by multiple speakers.

8. A method as claimed in claim 4, wherein determining whether the detected pattern of variability of the magnetic field is indicative of a reference pattern associated with playback of a predetermined spoken phrase through a loudspeaker comprises:

passing the detected pattern of variability of the magnetic field to a classifier that has been trained to with inputs obtained from playback of the predetermined spoken phrase through a loudspeaker.

9. A system for detecting a replay attack in a speaker recognition system, the system being configured for:

identifying at least one feature of a detected magnetic field;

determining whether the at least one identified feature of the detected magnetic field is indicative of playback of speech through a loudspeaker; and

if so, determining that a replay attack may have taken place,

wherein identifying at least one feature of the detected magnetic field comprises:

receiving a signal from a magnetometer; and

detecting modulation of the at least one feature from the received signal from the magnetometer at one or more of:

at least one frequency in the range of 2Hz-10Hz;

a speech syllabic rate; and/or

an articulation rate.

10. A system as claimed in claim 9, further configured for:

receiving an audio signal representing speech, wherein the audio signal is received at substantially the same time as the magnetic field is detected; and

if it is determined that the at least one identified feature of the detected magnetic field is indicative of playback of speech through a loudspeaker, determining that the audio signal may result from said replay attack.

11. A system as claimed in claim 9 or 10, configured for identifying at least one feature of the detected magnetic field by:

receiving a signal from a magnetometer; and

performing a Discrete Fourier Transform on the received signal from the magnetometer.

12. A system as claimed in claim 9, 10 or 11, configured for determining whether a detected pattern of variability of the magnetic field is indicative of a reference pattern associated with playback of a predetermined spoken phrase through a loudspeaker.

13. A system as claimed in claim 12, configured for determining whether the detected pattern of variability of the magnetic field is indicative of a reference pattern associated with playback of a predetermined spoken phrase through a loudspeaker by:

comparing the detected pattern of variability of the magnetic field with a stored reference pattern.

14. A system as claimed in claim 13, wherein the stored reference pattern corresponds to the predetermined spoken phrase, as spoken by a specific enrolled user.

15. A system as claimed in claim 13, wherein the stored reference pattern corresponds to the predetermined spoken phrase, as spoken by multiple speakers.

16. A system as claimed in claim 12, configured for determining whether the detected pattern of variability of the magnetic field is indicative of a reference pattern associated with playback of a predetermined spoken phrase through a loudspeaker by:

passing the detected pattern of variability of the magnetic field to a classifier that has been trained to with inputs obtained from playback of the predetermined spoken phrase through a loudspeaker.

17. A device comprising a system as claimed in any of claims 9 to 16.

18. A device as claimed in claim 17, wherein the device comprises a mobile telephone, an audio player, a video player, a mobile computing platform, a games device, a remote controller device, a toy, a machine, or a home automation controller or a domestic appliance.

19. A computer program product, comprising a computer-readable tangible medium, and instructions for performing a method according to any one of claims 1 to 8.

20. A non-transitory computer readable storage medium having computer-executable instructions stored thereon that, when executed by processor circuitry, cause the processor circuitry to perform a method according to any one of claims 1 to 8.

21. A device comprising the non-transitory computer readable storage medium as claimed in claim 20.

22. A device as claimed in claim 21, wherein the device comprises a mobile telephone, an audio player, a video player, a mobile computing platform, a games device, a remote controller device, a toy, a machine, or a home automation controller or a domestic appliance.

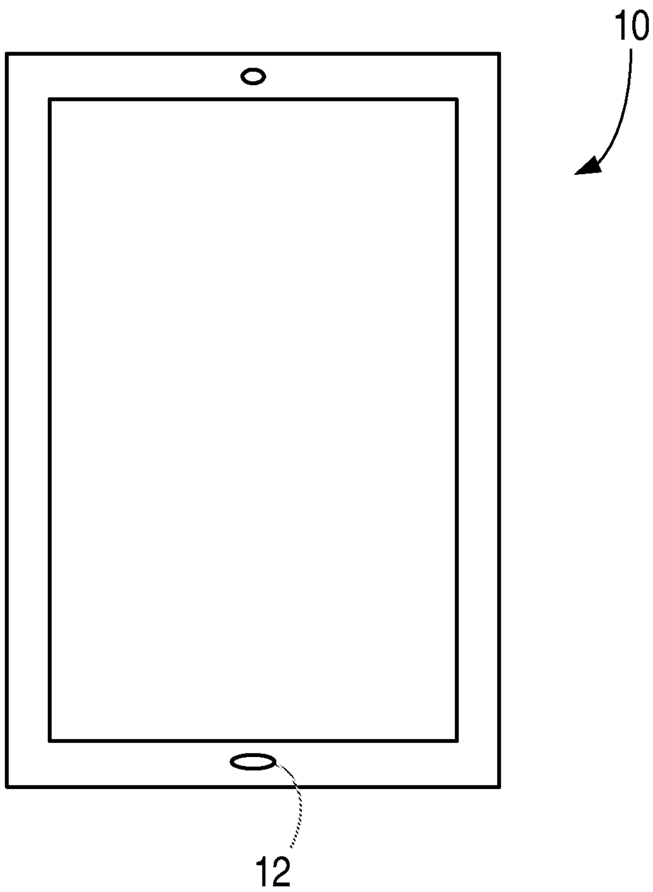

Figure 1

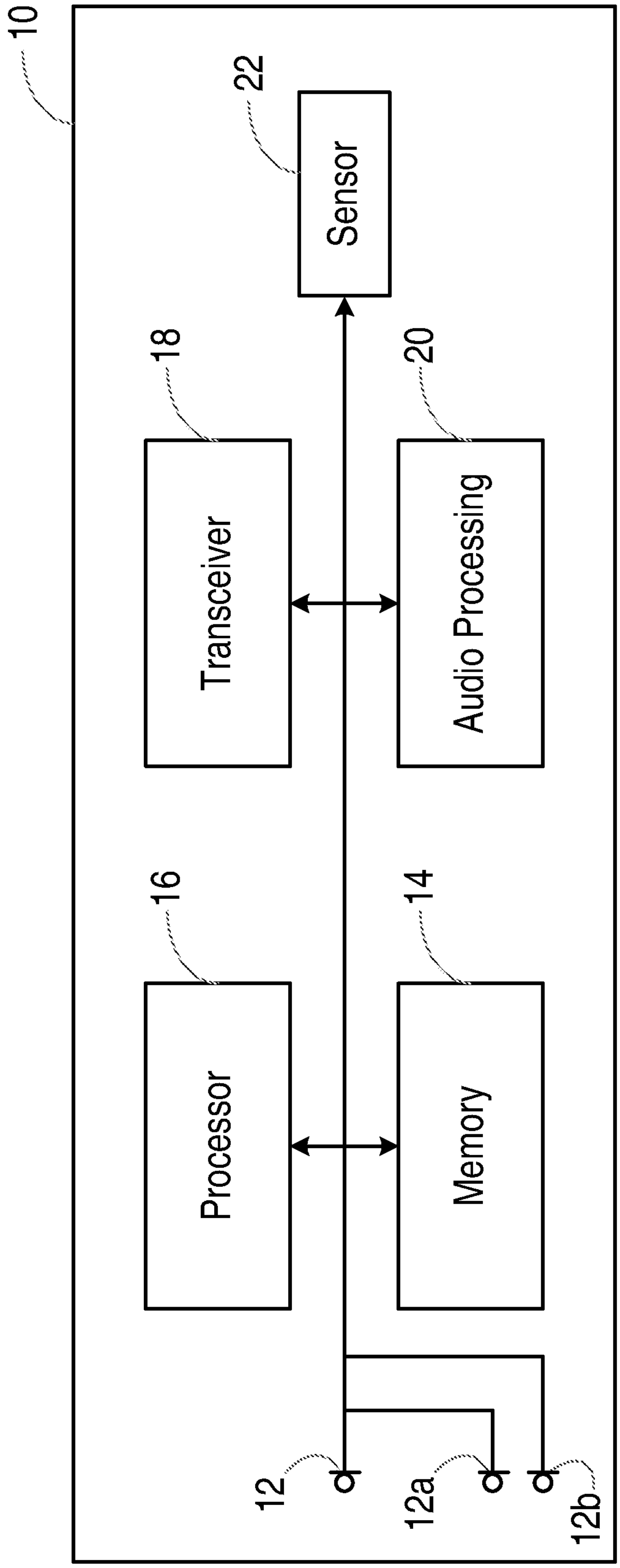
10
22
Sensor
18
Transceiver
20
Audio Processing
16
Processor
14
Memory
12
12a
12b

Figure 2

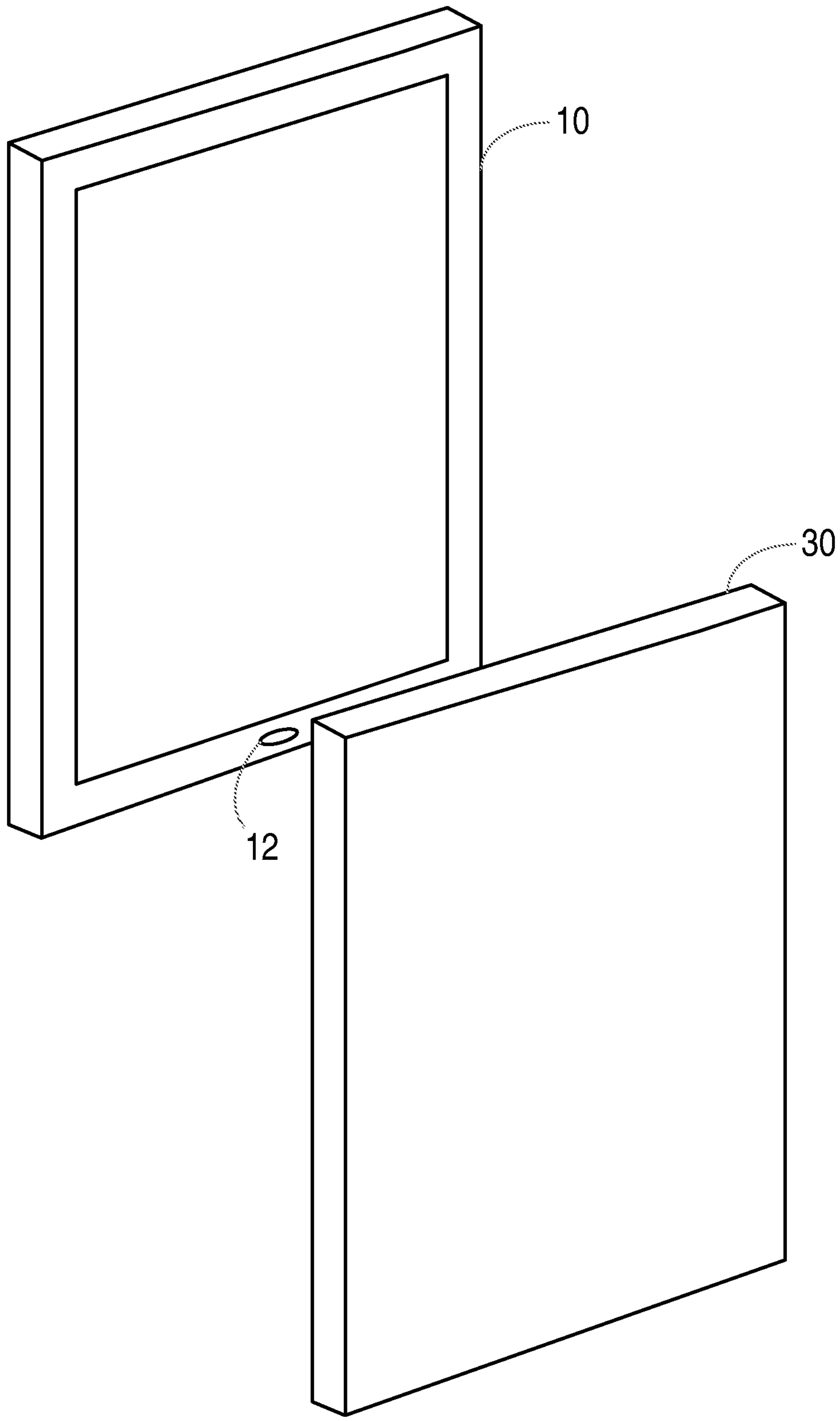

Figure 3

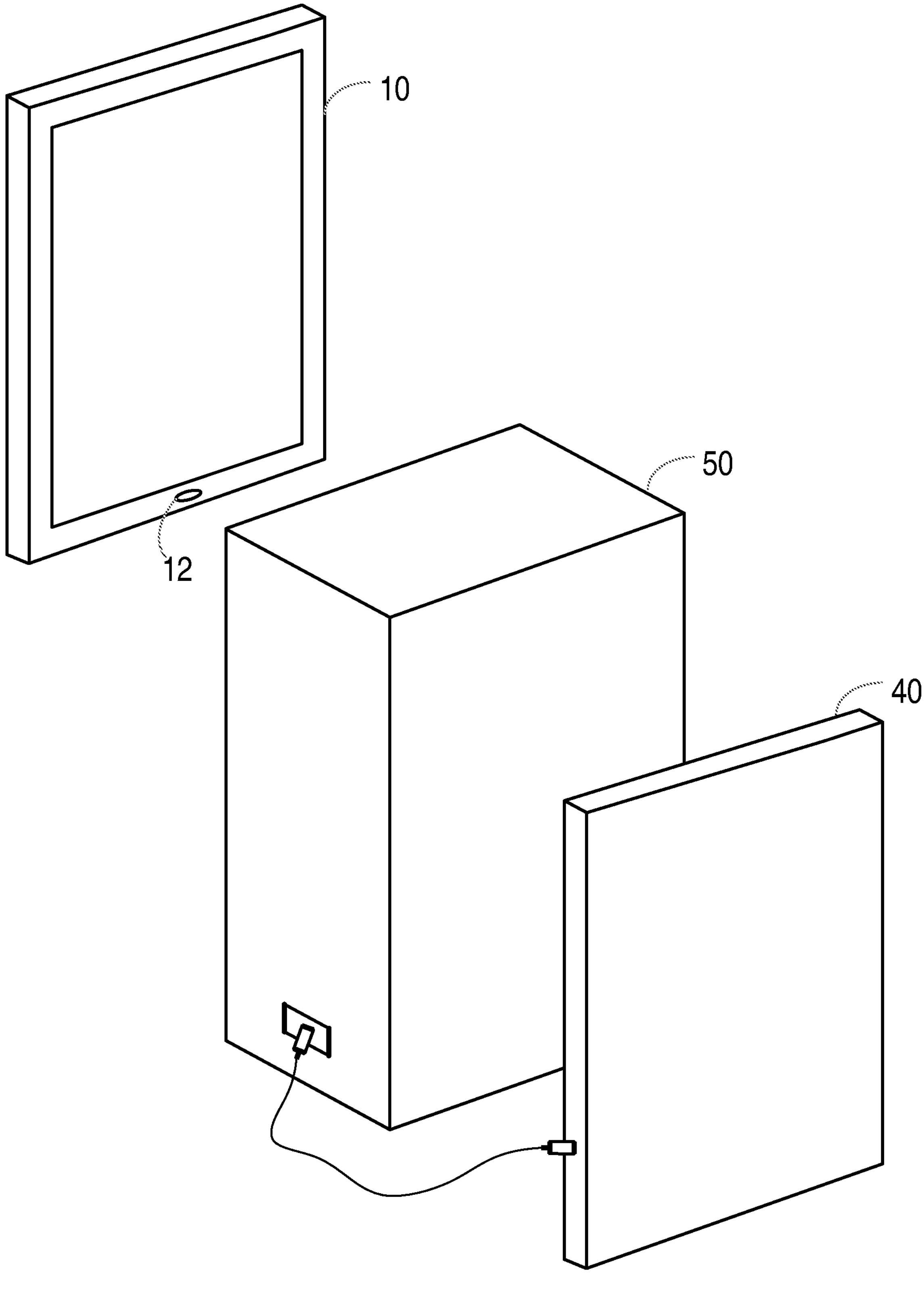

Figure 4

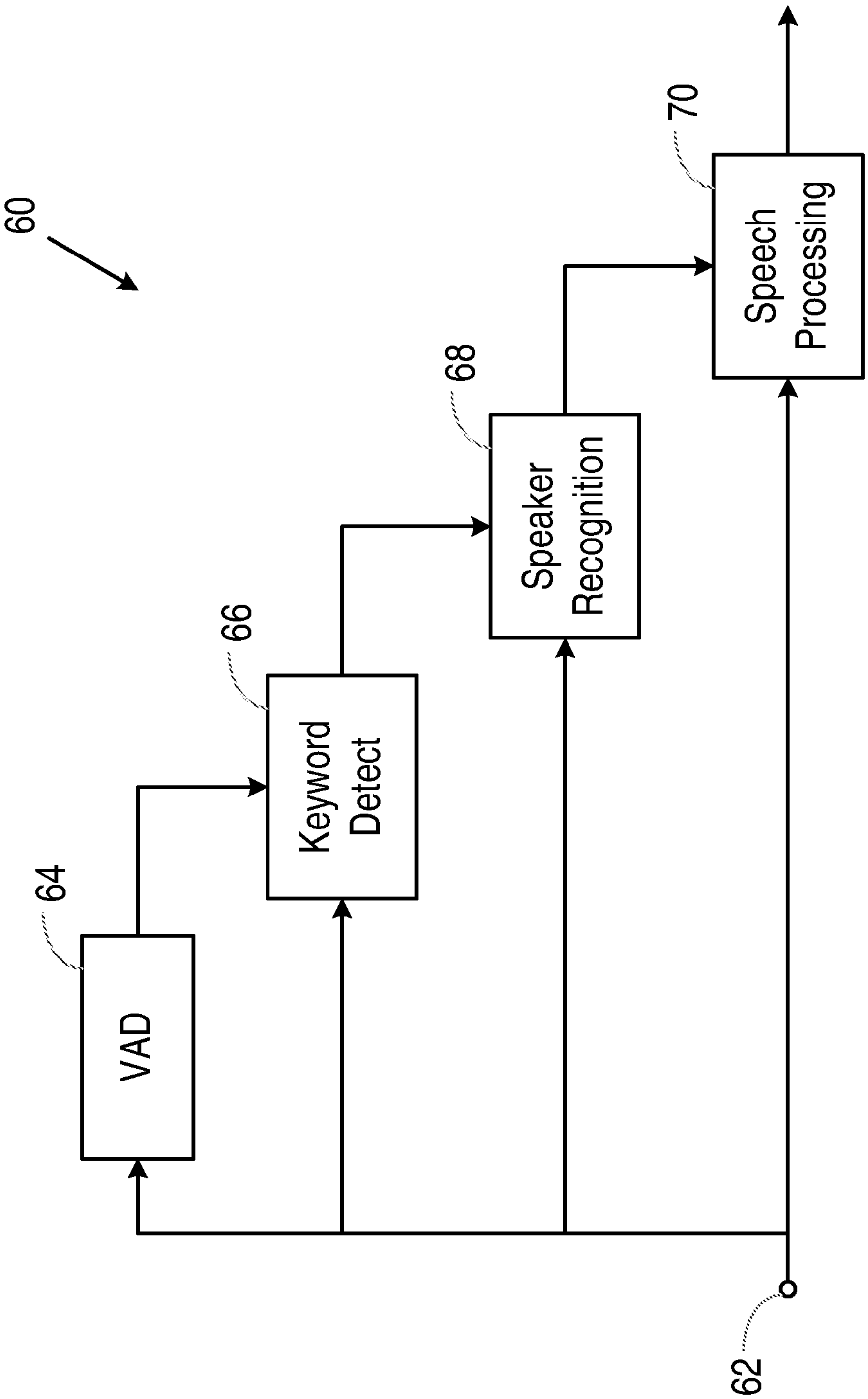
60
64
VAD
66
Keyword Detect
68
Speaker Recognition
70
Speech Processing
62

Figure 5

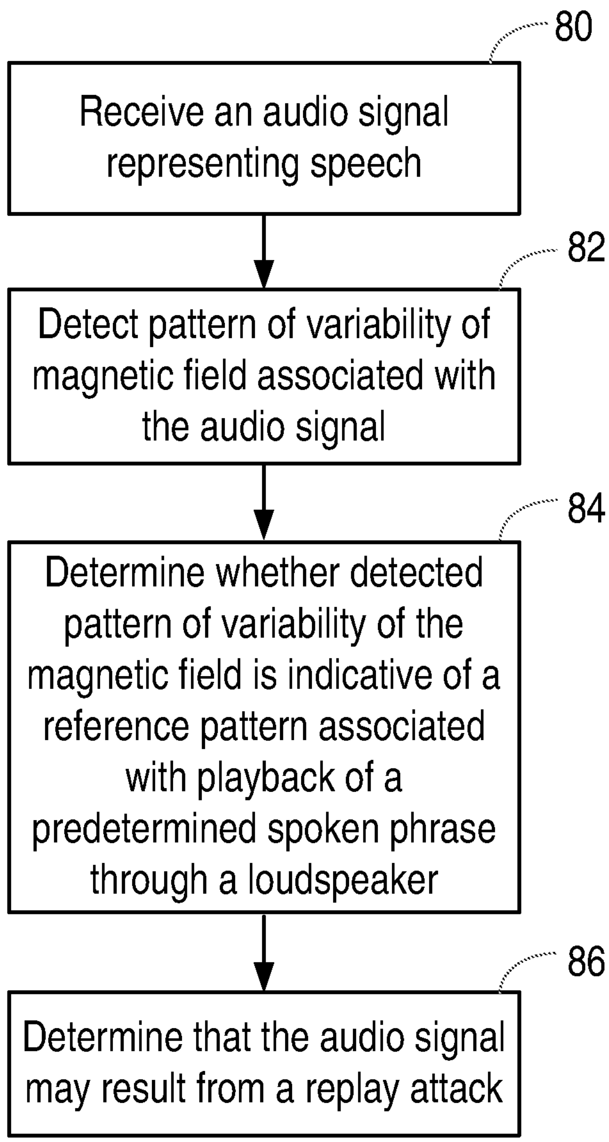
80
Receive an audio signal representing speech
82
Detect pattern of variability of magnetic field associated with the audio signal
84
Determine whether detected pattern of variability of the magnetic field is indicative of a reference pattern associated with playback of a predetermined spoken phrase through a loudspeaker
86
Determine that the audio signal may result from a replay attack

Figure 6

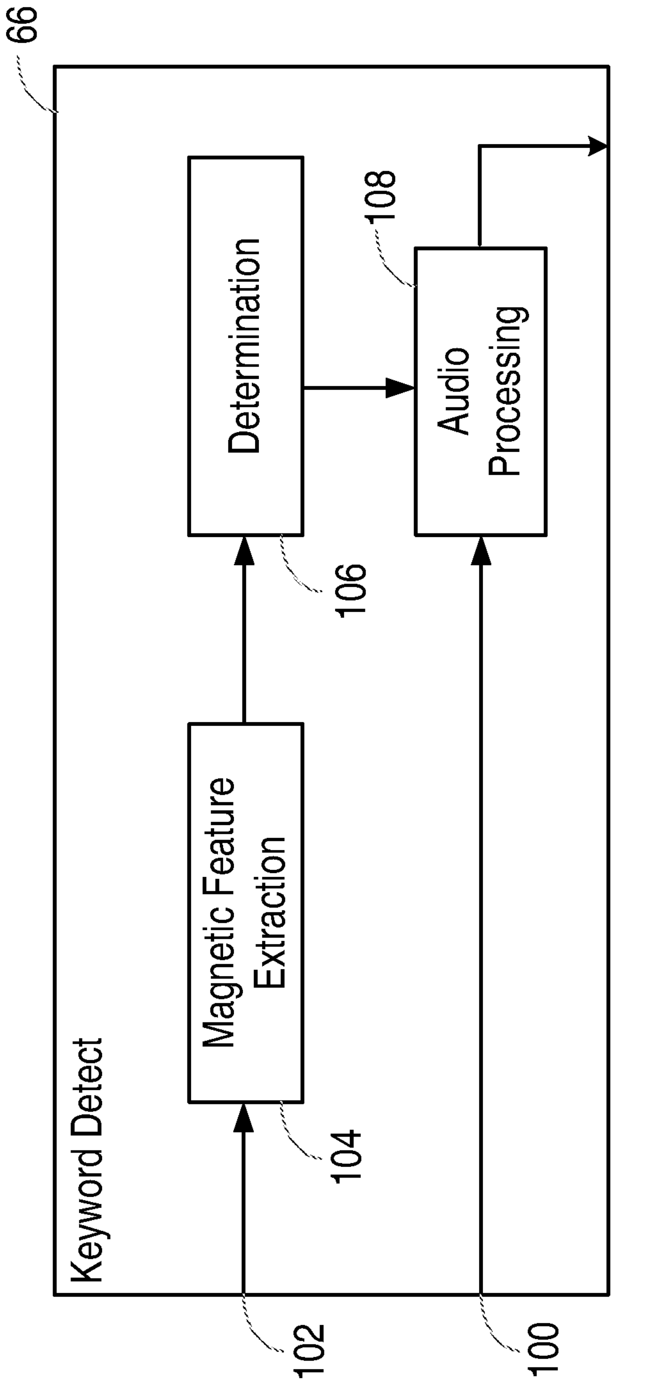
66
Keyword Detect
Magnetic Feature Extraction
Determination
Audio Processing
104
106
108
102
100

Figure 7

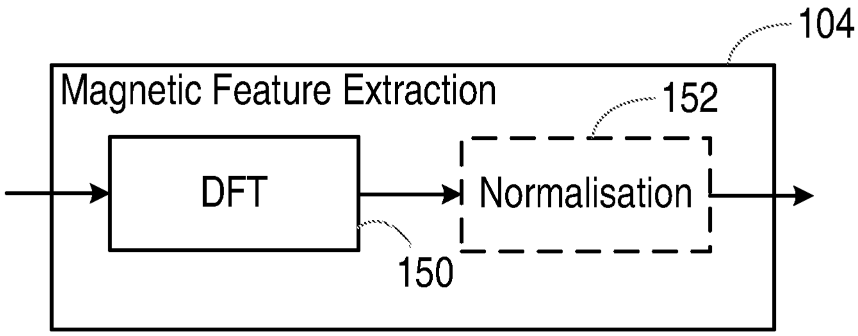

Figure 8

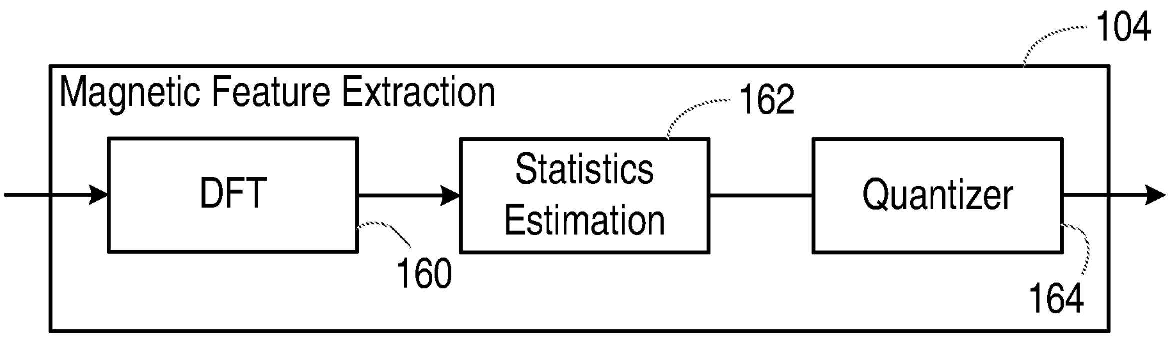

Figure 9

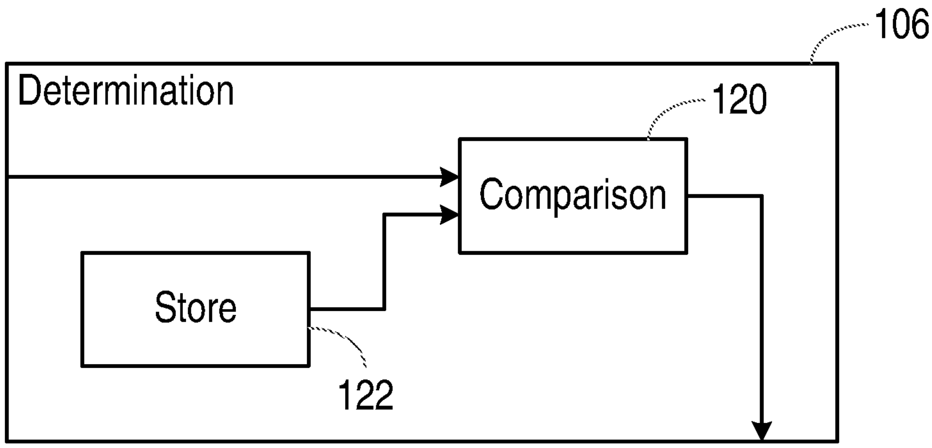

Figure 10

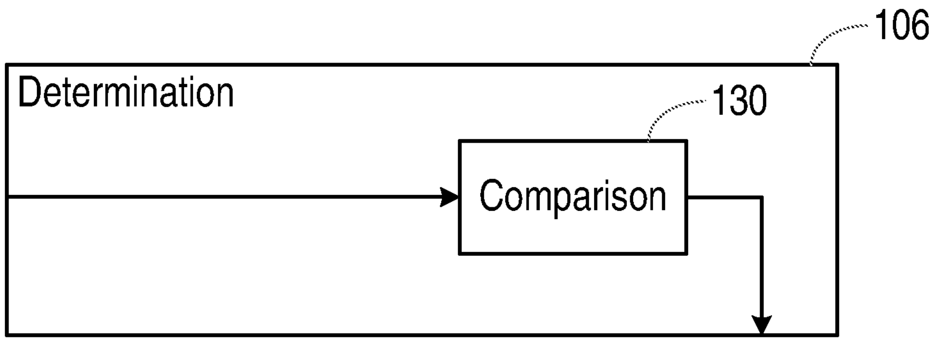

Figure 11

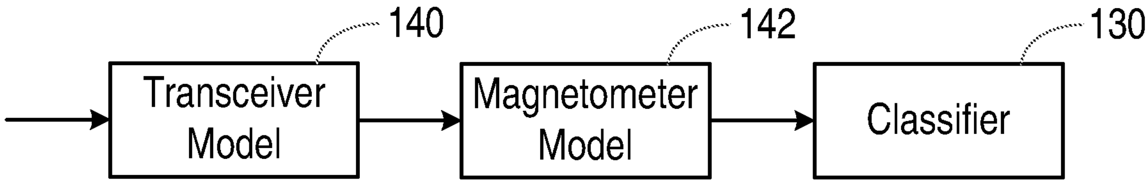

Figure 12

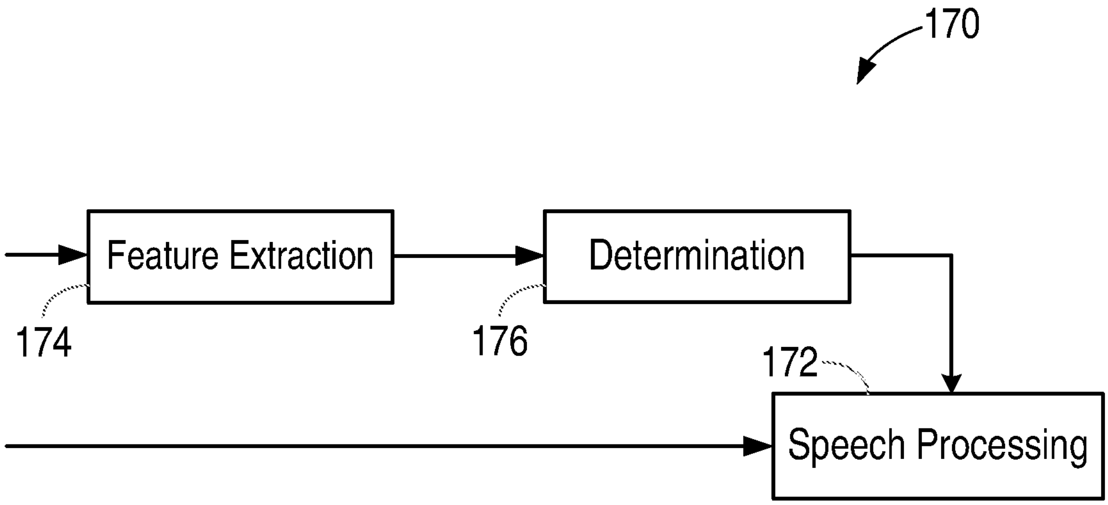

Figure 13

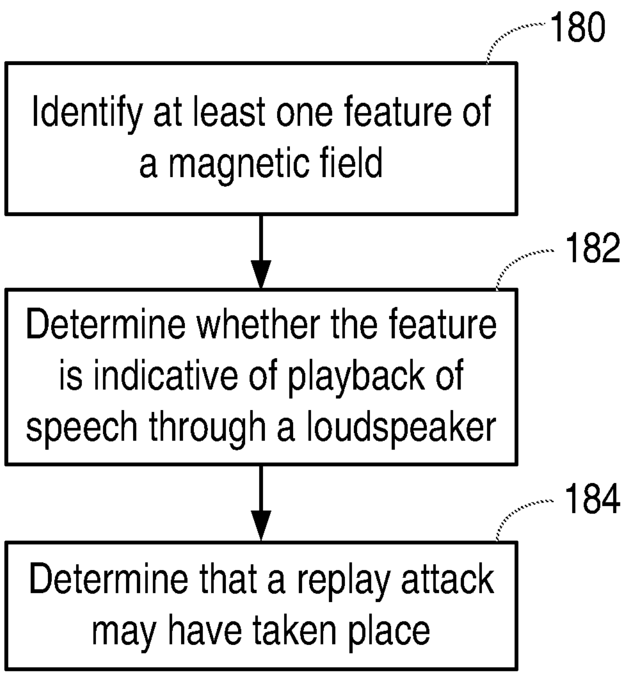

Figure 14

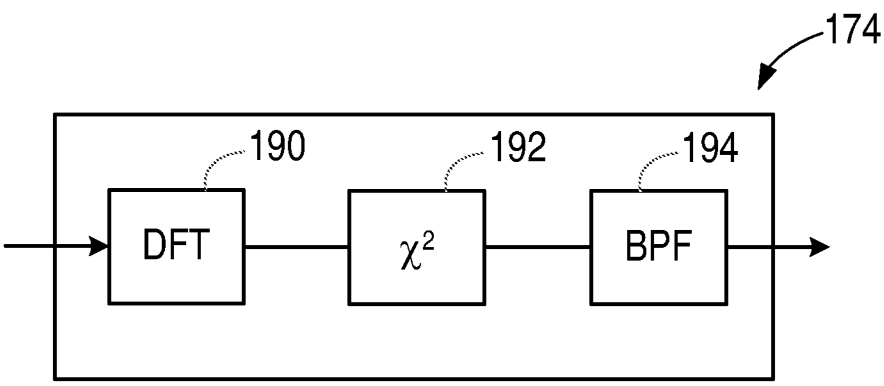

Figure 15

INTERNATIONAL SEARCH REPORT

International application No
PCT/GB2018/051787

A. CLASSIFICATION OF SUBJECT MATTER
INV. G10L17/02 G10L25/03 G06F21/32 H04L9/32 H04R29/00
H04W12/00
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G10L G06F H04L H04R H04W

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, COMPENDEX, INSPEC, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CHEN SI ET AL: "You Can Hear But You Cannot Steal: Defending Against Voice Impersonation Attacks on Smartphones", PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON DISTRIBUTED COMPUTING SYSTEMS, IEEE COMPUTER SOCIETY, US, 5 June 2017 (2017-06-05), pages 183-195, XP033122979, ISSN: 1063-6927, DOI: 10.1109/ICDCS.2017.133 [retrieved on 2017-07-13] figure 4 p. 185, section III. Problem Formulation, Section A. Adversary Model, Type 1: Voice Replay Attack section IV. The Proposed Solution p. 187, left-hand column, section A. System Architecture, second paragraph, line 2 from bottom	1-22

-/--

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
2 August 2018	16/08/2018

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Chétry, Nicolas

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
	p. 189, left-hand column, second paragraph, lines 1-5 from bottom	
A	JESUS VILLALBA ET AL: "Preventing replay attacks on speaker verification systems", SECURITY TECHNOLOGY (ICCST), 2011 IEEE INTERNATIONAL CARNAHAN CONFERENCE ON, IEEE, 18 October 2011 (2011-10-18), pages 1-8, XP032069118, DOI: 10.1109/CCST.2011.6095943 ISBN: 978-1-4577-0902-9 figures 4,5 section III. Far Field Replay Attack Detection System section III. section A. Features, third paragraph	1-22
A	US 7 016 833 B2 (UNIV CALIFORNIA [US]) 21 March 2006 (2006-03-21) column 3, lines 31-37	1-22
A	GB 2 493 849 A (BOEING CO [US]) 20 February 2013 (2013-02-20) figure 1 paragraphs [0004], [0012], [0020]	1-22

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/GB2018/051787

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 7016833 B2	21-03-2006	US 2005060153 A1	17-03-2005
		US 2007100608 A1	03-05-2007
GB 2493849 A	20-02-2013	GB 2493849 A	20-02-2013
		JP 6106373 B2	29-03-2017
		JP 2013045104 A	04-03-2013
		US 2013043977 A1	21-02-2013