



(51) International Patent Classification:

G10L 25/93 (2013.01) G10L 17/14 (2013.01)

(21) International Application Number:

PCT/GB2018/052905

(22) International Filing Date:

11 October 2018 (11.10.2018)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

62/571,978	13 October 2017 (13.10.2017)	US
62/578,667	30 October 2017 (30.10.2017)	US
1719731.0	28 November 2017 (28.11.2017)	GB
1719734.4	28 November 2017 (28.11.2017)	GB

(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 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: ANALYSING SPEECH SIGNALS

(57) Abstract: A method of analysis of an audio signal comprises: receiving an audio signal representing speech; extracting first and second components of the audio signal representing first and second acoustic classes of the speech respectively; analysing the first and second components of the audio signal with models of the first and second acoustic classes of the speech of an enrolled user. Based on the analysing, information is obtained about at least one of a channel and noise affecting the audio signal.

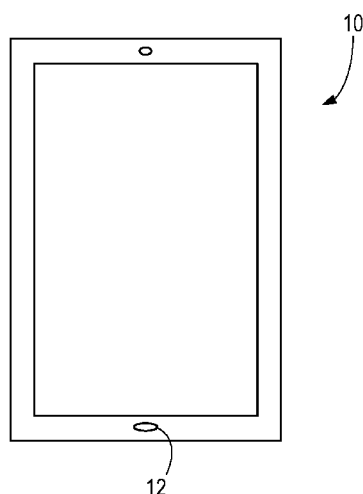


Figure 1

Published:

- *with international search report (Art. 21(3))*
- *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

ANALYSING SPEECH SIGNALS

Technical Field

- 5 Embodiments described herein relate to methods and devices for analysing speech signals.

Background

- 10 Many devices include microphones, which can be used to detect ambient sounds. In many situations, the ambient sounds include the speech of one or more nearby speaker. Audio signals generated by the microphones can be used in many ways. For example, audio signals representing speech can be used as the input to a speech recognition system, allowing a user to control a device or system using spoken
15 commands.

Summary

- According to a first aspect of the present invention, there is provided a method of
20 analysis of an audio signal, the method comprising: receiving an audio signal representing speech; extracting first and second components of the audio signal representing first and second acoustic classes of the speech respectively; analysing the first and second components of the audio signal with models of the first and second acoustic classes of the speech of an enrolled user; and based on said analysing,
25 obtaining information about at least one of a channel and noise affecting said audio signal.

- According to another aspect of the present invention, there is provided a system for analysing an audio signal, configured for performing the method.

30

- According to another aspect of the present invention, there is provided a device comprising such a system. 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
35 appliance.

According to another 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.

- 5 According to another 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.
- 10 According to a second aspect of the invention, there is provided a method of speaker identification, comprising: receiving an audio signal representing speech; removing effects of a channel and/or noise from the received audio signal to obtain a cleaned audio signal; obtaining an average spectrum of at least a part of the cleaned audio signal; comparing the average spectrum with a long term average speaker model for
15 an enrolled speaker; and determining based on the comparison whether the speech is the speech of the enrolled speaker.

- Obtaining an average spectrum of at least a part of the cleaned audio signal may comprise obtaining an average spectrum of a part of the cleaned audio signal
20 representing voiced speech.

- Obtaining an average spectrum of at least a part of the cleaned audio signal may comprise obtaining a first average spectrum of a part of the cleaned audio signal representing a first acoustic class and obtaining a second average spectrum of a part
25 of the cleaned audio signal representing a second acoustic class, and comparing the average spectrum with a long term average speaker model for an enrolled speaker may comprise comparing the first average spectrum with a long term average speaker model for the first acoustic class for the enrolled speaker and comparing the second average spectrum with a long term average speaker model for the second acoustic
30 class for the enrolled speaker.

The first acoustic class may be voiced speech and the second acoustic class unvoiced speech.

- 35 The method may comprise comparing the average spectrum with respective long term average speaker models for each of a plurality of enrolled speakers; and determining

based on the comparison whether the speech is the speech of one of the enrolled speakers.

5 The method may further comprise comparing the average spectrum with a Universal Background Model; and including a result of the comparing the average spectrum with the Universal Background Model in determining whether the speech is the speech of one of the enrolled speakers.

10 The method may comprise identifying one of the enrolled speakers as a most likely candidate as a source of the speech.

15 The method may comprise: obtaining information about the effects of a channel and/or noise on the received audio signal by: receiving the audio signal representing speech; extracting first and second components of the audio signal representing first and second acoustic classes of the speech respectively; analysing the first and second components of the audio signal with models of the first and second acoustic classes of the speech of an enrolled user; and, based on said analysing, obtaining information about at least one of a channel and noise affecting said audio signal.

20 The method may comprise analysing the first and second components of the audio signal with models of the first and second acoustic classes of the speech of a plurality of enrolled users, to obtain respective hypothetical values of the channel, and determining that the speech is not the speech of any enrolled speaker whose models give rise to physically implausible hypothetical values of the channel.

25 A hypothetical value of the channel may be considered to be physically implausible if it contains variations exceeding a threshold level across the relevant frequency range. A hypothetical value of the channel may be considered to be physically implausible if it contains significant discontinuities.

30 According to another aspect of the present invention, there is provided a system for analysing an audio signal, configured for performing the method.

35 According to another aspect of the present invention, there is provided a device comprising such a system. 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.

5 According to another 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 second aspect.

10 According to another 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 second aspect.

Brief Description of Drawings

15

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.

20

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

Figure 3 is a flow chart illustrating a method of analysing an audio signal;

25 Figure 4 is a block diagram illustrating a system for analysing an audio signal;

Figure 5 illustrates results in the method of Figure 3;

30 Figure 6 is a block diagram illustrating an alternative system for analysing an audio signal;

Figure 7 is a block diagram illustrating a further alternative system for analysing an audio signal;

35 Figure 8 is a block diagram illustrating a further alternative system for analysing an audio signal;

Figure 9 illustrates a possible relay attack on a voice biometric system;

Figure 10 illustrates an effect of a replay attack;

5

Figure 11 is a flow chart illustrating a method of detecting a replay attack;

Figure 12 is a flow chart, illustrating a method of identifying a speaker;

10 Figure 13 is a block diagram illustrating a system for identifying a speaker; and

Figure 14 is a block diagram illustrating a system for identifying a speaker.

15 **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

20 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.

The methods described herein can be implemented in a wide range of devices and

25 systems. However, for ease of explanation of one embodiment, an illustrative example will be described, in which the implementation occurs in a smartphone.

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
30 user who is holding the smartphone 10 close to their face.

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

Specifically, Figure 2 shows various interconnected components of the smartphone 10.

35 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,
5 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.

10

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.

15 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.

20 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.

25 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
30 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
35 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
5 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
10 returned to the smartphone 10 or other local device.

Methods described herein proceed from the recognition that different parts of a user's speech have different properties.

15 Specifically, it is known that speech can be divided into voiced sounds and unvoiced or voiceless sounds. A voiced sound is one in which the vocal cords of the speaker vibrate, and a voiceless sound is one in which they do not.

It is now recognised that the voiced and unvoiced sounds have different frequency
20 properties, and that these different frequency properties can be used to obtain useful information about the speech signal.

Figure 3 is a flow chart, illustrating a method of analysing an audio signal, and **Figure**
25 **4** is a block diagram illustrating functional blocks in the analysis system.

Specifically, in step 50 in the method of Figure 3, an audio signal, which is expected to contain speech, is received on an input 70 of the system shown in Figure 4.

30 The received signal is divided into frames, which may for example have lengths in the range of 10-100 ms, and then passed to a voiced/unvoiced detection block 72. Thus, in step 52 of the process, first and second components of the audio signal, representing different first and second acoustic classes of the speech, are extracted from the received signal. Extracting the first and second components of the audio
35 signal may comprise identifying periods when the audio signal contains the first acoustic class of speech, and identifying periods when the audio signal contains the

second acoustic class of speech. More specifically, extracting the first and second components of the audio signal may comprise identifying frames of the audio signal that contain the first acoustic class of speech, and frames that contain the second acoustic class of speech.

5

When the first and second acoustic classes of the speech are voiced speech and unvoiced speech, there are several methods that can be used to identify voiced and unvoiced speech, for example: using a deep neural network (DNN), trained against a golden reference, for example using Praat software; performing an autocorrelation with unit delay on the speech signal (because voiced speech has a higher autocorrelation for non-zero lags); performing a linear predictive coding (LPC) analysis (because the initial reflection coefficient is a good indicator of voiced speech); looking at the zero-crossing rate of the speech signal (because unvoiced speech has a higher zero-crossing rate); looking at the short term energy of the signal (which tends to be higher for voiced speech); tracking the first formant frequency F0 (because unvoiced speech does not contain the first format frequency); examining the error in a linear predictive coding (LPC) analysis (because the LPC prediction error is lower for voiced speech); using automatic speech recognition to identify the words being spoken and hence the division of the speech into voiced and unvoiced speech; or fusing any or all of the above.

20

In the embodiments described further below, the first and second acoustic classes of the speech are voiced speech and unvoiced speech. However, the first and second acoustic classes of the speech may be any phonetically distinguishable acoustic classes. For example, they may be different phoneme classes, for example two different sets of vowels; they may be two different fricatives; or the first class may be fricatives while the second class are sibilants.

25

The received signal may be supplied to a voice activity detection block, and only supplied to the voiced/unvoiced detection block 72 when it is determined that it does contain speech. In that case, or otherwise when there is reason to believe that the audio signal contains only speech, the step of identifying periods when the audio signal contains unvoiced speech may comprise identifying periods when the audio signal contains voiced speech, and identifying the remaining periods of speech as containing unvoiced speech.

35

The voiced/unvoiced detection block 72 may for example be based on Praat speech analysis software.

5 The voiced/unvoiced detection block 72 thus outputs the first component of the audio signal, S_v , representing voiced speech and the second component, S_u , representing unvoiced speech.

10 More specifically, in some embodiments, the first component of the audio signal, S_v , representing voiced speech and the second component, S_u , representing unvoiced speech, are averaged spectra of the voiced and unvoiced components of the speech. By averaged spectra are meant spectra of the speech obtained and averaged over multiple frames.

15 The spectra can be averaged over enough data to provide reasonable confidence in the information that is obtained about the speech signal. In general terms, this information will become more reliable as more data is used to form the average spectra.

20 In some cases, spectra averaged over 500ms of the relevant speech will be enough to provide reliable averaged spectra. The length of time over which the averaged spectra are generated may be adapted based on the articulation rate of the speech, in order to ensure that the speech contains enough phonetic variation to provide a reliable average. The length of time over which the averaged spectra are generated may be adapted based on the content of the speech. If the user is speaking a predetermined
25 known phrase, this may be more discriminative than speaking words of the user's choosing, and so a useful average can be obtained in a shorter period. The process illustrated in Figure 3 may be performed regularly while the user is speaking, providing regularly updated information at the end of the method as more speech is received. It may then be judged that enough speech has been processed when the results of the
30 method converge to stable values.

The signal received on the input 70 is also passed to a speaker recognition block 74, which performs a voice biometric process to identify the speaker, from amongst a plurality of enrolled speakers. The process of enrolment in a speaker recognition
35 system typically involves the speaker providing a sample of speech, from which specific features are extracted, and the extracted features are used to form a model of

the speaker's speech. In use, corresponding features are extracted from a sample of speech, and these are compared with the previously obtained model to obtain a measure of the likelihood that the speaker is the previously enrolled speaker.

- 5 In some situations, the speaker recognition system attempts to identify one or more enrolled speaker without any prior expectation as to who the speaker should be. In other situations, there is a prior expectation as to who the speaker should be, for example because there is only one enrolled user of the particular device that is being used, or because the user has already identified themselves in some other way.

10

In this illustrated example, the speaker recognition block 74 is used to identify the speaker. In other examples, there may be an assumption that the speaker is a particular person, or is selected from a small group of people.

- 15 In step 54 of the process shown in Figure 3, the first and second components of the audio signal are compared with models of the first acoustic class (for example the voiced component) of the speech of an enrolled user and of the second acoustic class (for example the unvoiced component) of the speech of the enrolled user. For example, comparing the first and second components of the audio signal with the
20 models of the voiced and unvoiced speech of the enrolled user may comprise comparing magnitudes of the audio signal at a number of predetermined frequencies with magnitudes in the models

- Thus, in the system shown in Figure 4, one or more speaker model is stored, for example in a database. Based on the output of the speaker recognition block 74, or
25 based on a prior assumption as to who the speaker is expected to be, one or more speaker model is selected.

- In this embodiment, each speaker model contains separate models of the voiced speech and the unvoiced speech of the enrolled user. More specifically, the model of
30 the voiced speech and the model of the unvoiced speech of the enrolled user each comprise amplitude values corresponding to multiple frequencies.

Thus, **Figure 5** shows a multiple speaker models.

- 35 Specifically, each speaker model shown in Figure 5 comprises a long term averaged spectrum of the voiced components of the speech and a long term averaged spectrum

of the unvoiced components of the speech. These models are obtained from the respective speakers during previous separate enrolment processes, during which the speakers speak, either uttering predetermined standard test phrases or saying words of their own choosing.

5

Figure 5 shows the speaker models for five speakers, labelled Speaker 1 – Speaker 5. The model for Speaker 1 comprises the long term averaged spectrum 90 of the voiced components of the speech and the long term averaged spectrum 91 of the unvoiced components of the speech; the model for Speaker 2 comprises the long term averaged spectrum 92 of the voiced components of the speech and the long term averaged spectrum 93 of the unvoiced components of the speech; the model for Speaker 3 comprises the long term averaged spectrum 94 of the voiced components of the speech and the long term averaged spectrum 95 of the unvoiced components of the speech; the model for Speaker 4 comprises the long term averaged spectrum 96 of the voiced components of the speech and the long term averaged spectrum 97 of the unvoiced components of the speech; and the model for Speaker 5 comprises the long term averaged spectrum 98 of the voiced components of the speech and the long term averaged spectrum 99 of the unvoiced components of the speech.

20 In each case, the model of the speech comprises a vector containing amplitude values at a plurality of frequencies.

The plurality of frequencies may be selected from within a frequency range that contains the most useful information for discriminating between speakers. For example, the range may be from 20Hz to 8kHz, or from 20Hz to 4kHz.

The frequencies at which the amplitude values are taken may be linearly spaced, with equal frequency spacings between each adjacent pair of frequencies. Alternatively, the frequencies may be non-linearly spaced. For example, the frequencies may be equally spaced on the mel scale.

The number of amplitude values used to form the model of the speech may be chosen depending on the frequency spacings. For example, using linear spacings the model may contain amplitude values for 64 to 512 frequencies. Using mel spacings, it may be possible to use fewer frequencies, for example between 10 and 20 mel-spaced frequencies.

Thus, the model of the voiced speech may be indicated as M_v , where M_v represents a vector comprising one amplitude value at each of the selected frequencies, while the model of the unvoiced speech may be indicated as M_u , where M_u represents a
 5 vector comprising one amplitude value at each of the selected frequencies.

As will be appreciated, the received signal, containing the user's speech, will be affected by the properties of the channel, which we take to mean any factor that produces a difference between the user's speech and the speech signal as generated
 10 by the microphone alters, and the received signal will also be affected by noise.

Thus, assuming that the channel and the noise are constant over the period during which the received signal is averaged to form the first and second components of the received speech, these first and second components can be expressed as:

15

$$S_v = \alpha M_v + n, \text{ and}$$

$$S_u = \alpha M_u + n,$$

where

20 α represents the frequency spectrum of a multiplicative disturbance component, referred to herein as the channel, and
 n represents the frequency spectrum of an additive disturbance component, referred to herein as the noise.

25 Thus, with measurements S_v and S_u , and with models M_v and M_u , these two equations can therefore be solved for the two unknowns, α and n .

Thus, for illustrative purposes,

$$\alpha = \frac{(S_u - S_v)}{(M_u - M_v)}, \text{ and}$$

30
$$n = \frac{(S_u M_v - S_v M_u)}{(M_u - M_v)}.$$

For completeness, it should be noted that, with measurements of the spectrum made at a plurality of frequencies, these two equations are effectively solved at each of the frequencies.

- 5 Alternatively, with measurements made at f different frequencies, the equations $S_v = \alpha M_v + n$, and $S_u = \alpha M_u + n$ can each be regarded as f different equations to be solved.

10 In that case, having solved the equations, it may be useful to apply a low-pass filter, or a statistical filter such as a Savitsky-Golay filter, to the results in order to obtain low-pass filtered versions of the channel and noise characteristics.

As an alternative example, a least squares method may be used to obtain solutions to the $2f$ different equations.

15

It will be noted that the calculations set out above rely on determining the difference ($M_u - M_v$) between the model of the unvoiced speech and the model of the voiced speech. Where these are similar, for example in the range 1.3 – 1.6kHz in the case of Speaker 1 in Figure 5, then any small uncertainties in either of the models will
20 potentially be magnified into large errors in the calculated values for the channel and/or the noise. Thus, the calculated values in any such frequency ranges may be given lower significance in any subsequent processing steps that use the calculated values, for example a reduced can weight applied to the values used in later processing steps. Alternatively, when it is known in advance that the model of the unvoiced speech and
25 the model of the voiced speech are similar in a particular frequency range, the equations given above need not be solved for frequencies in this range.

Thus, as shown at step 56 of the process shown in Figure 3, information is obtained about the channel and/or the noise affecting the audio signal.

30

This information can be used in many different ways.

Figure 6 illustrates one such use. The system shown in Figure 6 is similar to the system of Figure 4, and the same reference numerals are used to refer to the same
35 components of the system.

In the system of Figure 6, the comparison block 78 is used to obtain information about the channel α that is affecting the received audio signal. Specifically, the comparison block 78 may be used to obtain the frequency spectrum of the channel. This can be used to compensate the received audio signal to take account of the channel.

5

For one example, Figure 6 shows a channel compensation block 120, to which the audio signal received on the input 70 is supplied. The channel compensation block 120 also receives the frequency spectrum of the channel α . The channel compensation block 120 acts to remove the effects of the channel from the received signal, by dividing the received signal by the calculated channel α , before the received signal is passed to the speaker recognition block 74.

Thus, the output of the speaker recognition block 74, on the output 122, can be improved. That is, it can provide more reliable information about the identity of the speaker. This can then be supplied to a processing block 124 and used for any required purposes.

The output of the channel compensation block 120, containing the received signal after the effects of the channel have been removed, can be supplied to any suitable processing block 126, such as a speech recognition system, or the like.

Figure 7 illustrates another such use. The system shown in Figure 7 is similar to the system of Figure 4, and the same reference numerals are used to refer to the same components of the system.

25

In the system of Figure 7, the comparison block 78 is used to obtain information about the noise n that is affecting the received audio signal. Specifically, the comparison block 78 may be used to obtain the frequency spectrum of the noise. This can be used to take account of the noise when processing the received audio signal.

30

For one example, Figure 7 shows a filter block 128, to which the audio signal received on the input 70 is supplied. The filter block 128 also receives the frequency spectrum of the noise n . The filter block 128 acts so as to ensure that noise does not adversely affect the operation of the speaker recognition block 74.

35

For example, the calculated noise characteristic, n , can be subtracted from the received signal before any further processing takes place.

5 In another example, where the level of noise exceeds a predetermined threshold level at one or more frequencies, such that the operation of the speaker recognition block 74 could be compromised, the filter block 128 can remove the corrupted components of the received audio signal at those frequencies, before passing the signal to the speaker recognition block 74. Alternatively, these components could instead be flagged as being potentially corrupted, before being passed to the speaker recognition block 74 or
10 any further signal processing block.

Thus, the output of the speaker recognition block 74, on the output 122, can be improved. That is, it can provide more reliable information about the identity of the speaker. This can then be supplied to any suitable processing block 124, and used for
15 any required purposes.

The output of the filter block 128, containing the received signal after the frequency components that are excessively corrupted by noise have been removed, can be supplied to any suitable processing block 130, such as a speech recognition system, or
20 the like.

Figure 8 illustrates another such use. The system shown in Figure 8 is similar to the system of Figure 4, and the same reference numerals are used to refer to the same components of the system.

25

In the system of Figure 8, the comparison block 78 is used to obtain information about the channel α and the noise n that are affecting the received audio signal. Specifically, the comparison block 78 may be used to obtain the frequency spectrum of the channel and of the noise. This can be used to take account of the channel and the noise when
30 processing the received audio signal.

For one example, Figure 8 shows a combined filter block 134, to which the audio signal received on the input 70 is supplied. The combined filter block 134 also receives the frequency spectrum of the channel α and the noise n . The combined filter block 134
35 acts so as to ensure that channel effects and noise do not adversely affect the operation of the speaker recognition block 74.

For example, the calculated noise characteristic, n , can be subtracted from the received signal, and the remaining signal can be divided by the calculated channel α , before any further processing takes place.

5

Thus, the output of the speaker recognition block 74, on the output 122, can be improved. That is, it can provide more reliable information about the identity of the speaker. This can then be supplied to any suitable processing block 124, and used for any required purposes.

10

The output of the combined filter block 134, containing the received signal after the effects of the channel and the noise have been removed, can be supplied to any suitable processing block 136, such as a speech recognition system, or the like.

15

A further use of the information obtained about the channel and/or the noise affecting the audio signal is to overcome an attempt to deceive a voice biometric system by playing a recording of an enrolled user's voice in a so-called replay or spoof attack.

20 Additionally, a further use of the information obtained about the channel and/or the noise affecting the audio signal is to remove their effects from a received audio signal, meaning that the average spectrum of the speech contained in the audio signal can be used as a biometric.

25 **Figure 9** shows an example of a situation in which a replay attack is being performed. Thus, in Figure 9, 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 determine that the enrolled user's voice that it recognises is a recording, the attacker will gain access to one or more services that are intended to be accessible only by the enrolled user.

35

It is known that smartphones, such as the smartphone 30, are typically provided with loudspeakers that are of relatively low quality. 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.

5

Figure 10 illustrates the frequency response of a typical loudspeaker. Thus, at frequencies below a lower threshold frequency f_L , the loudspeaker suffers from low-frequency roll-off, as the bass response is limited by the size of the loudspeaker diaphragm. At frequencies above an upper threshold frequency f_U , the loudspeaker suffers from high-frequency roll-off. At frequencies between the lower threshold frequency f_L and the upper threshold frequency f_U , there is a degree of pass-band ripple, as the magnitude of the response varies periodically between β_1 and β_2 .

The size of these effects will be determined by the quality of the loudspeaker. For example, in a high quality loudspeaker, the lower threshold frequency f_L and the upper threshold frequency f_U should be such that there is minimal low-frequency roll-off or high-frequency roll-off within the frequency range that is typically audible to humans. However, size and cost constraints mean that many commercially available loudspeakers, such as those provided in smartphones such as the smartphone 30, do suffer from these effects to some extent.

Similarly, the magnitude of the pass-band ripple, that is the difference between β_1 and β_2 , will also depend on the quality of the loudspeaker.

If the voice of a speaker is played back through a loudspeaker whose frequency response has the general form shown in Figure 10, then this may be detectable in the received audio signal containing the speech of that speaker. It has previously been recognised that, if a received audio signal has particular frequency characteristics, that may be a sign that the received audio signal is the result of a replay attack. However, the frequency characteristics of the received signal depend on other factors, such as the frequency characteristics of the speech itself, and the properties of any ambient noise, and so it is difficult to make a precise determination that a signal comes from a replay attack based only on the frequency characteristics of the received signal.

However, the method shown in Figure 3, and described with reference thereto, can be used to make a more reliable determination as to whether a signal comes from a replay attack.

- 5 In one possibility, as shown in Figure 7, the frequency characteristic of the ambient noise is determined, and this is subtracted from the received audio signal by means of the filter 128. The received signal, with noise removed, is supplied to a processing block 130, which in this case may be a replay attack detection block.
- 10 For example, the replay attack detection block may perform any of the methods disclosed in EP-2860706A, such as testing whether a particular spectral ratio (for example a ratio of the signal energy from 0-2kHz to the signal energy from 2-4kHz) has a value that may be indicative of replay through a loudspeaker, or whether the ratio of the energy within a certain frequency band to the energy of the complete frequency
- 15 spectrum has a value that may be indicative of replay through a loudspeaker.

In another possibility, the method shown in Figure 3 is used to determine the frequency characteristic of the channel that affects the received speech. If the speech has been played back through a loudspeaker, the frequency response of the loudspeaker should

20 be visible in the frequency characteristic of the channel.

Figure 11 is a flow chart, illustrating a method of determining whether the received signal may result from a replay attack.

- 25 In the method of Figure 11, in step 140, an audio signal is received, representing speech.

In step 142, information is obtained about a channel affecting said audio signal. For example, the information about the channel may be obtained by the method shown in

30 Figure 3.

In step 144, it is determined whether the channel has at least one characteristic of a loudspeaker.

- 35 As shown at step 146, determining whether the channel has at least one characteristic of a loudspeaker may comprise determining whether the channel has a low frequency

roll-off. For example, the low-frequency roll-off may involve the measured channel decreasing at a relatively constant rate, such as 6dB per octave, for frequencies below a lower cut-off frequency f_L , which may for example be in the range 50Hz – 700Hz.

- 5 As shown at step 148, determining whether the channel has at least one characteristic of a loudspeaker may comprise determining whether the channel has a high frequency roll-off. For example, the high-frequency roll-off may involve the measured channel decreasing at a relatively constant rate, such as 6dB per octave, for frequencies above an upper cut-off frequency f_U , which may for example be in the range 18kHz – 24kHz.

10

- As shown at step 150, determining whether the channel has at least one characteristic of a loudspeaker may comprise determining whether the channel has ripple in a pass-band thereof. For example, this may comprise applying a Welch periodogram to the channel, and determining whether there is a predetermined amount of ripple in the characteristic. A degree of ripple (that is, a difference between β_1 and β_2 in the frequency response shown in Figure 10) exceeding a threshold value, such as 1dB, and with a peak-to-trough frequency of about 100Hz, over the central part of the pass-band, for example from 100Hz – 10kHz, can be regarded as characteristic of a loudspeaker.

20

For example, two or three of the steps 146, 148 and 150 may be performed, with the results being applied to a classifier, to determine whether the results of those steps are indeed characteristic of a loudspeaker frequency response.

- 25 As a further example, the channel frequency response can be applied as an input to a neural network, which has been trained to distinguish channels that are characteristic of loudspeakers from other channels.

- If it is determined that the channel has a characteristic of a loudspeaker, then it may be determined, perhaps on the basis of other indicators too, that the received audio signal is the result of a replay attack. In that case, the speech in the received audio signal may be disregarded when attempting to verify that the speaker is the expected enrolled speaker.

- 35 **Figure 12** is a flow chart, illustrating a method of speaker identification, and **Figure 13** is a block diagram of a system for performing speaker identification.

As described above, the system may be implemented in a smartphone, such as the smartphone 10, or any other device with voice biometric 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.

In step 160 of the method of Figure 12, the signal generated by a microphone 12 in response to ambient sound is received.

The received signal is divided into frames, which may for example have lengths in the range of 10-100 ms. These frames can be analysed to determine whether they represent speech, and only frames that represent speech are considered further.

The frames that represent speech are passed to a channel/noise removal block 180 and, in step 162 of the method, the effects of a channel and/or noise are removed from the received audio signal to obtain a cleaned audio signal.

The effects of the channel and/or noise can be determined by the method described above, or by any other suitable method, leaving a cleaned audio signal that is not adversely affected by any channel or noise effects.

- 5 In step 164 of the method, the cleaned audio signal is passed to an averaging block 182, which obtains an average spectrum of at least a part of the cleaned audio signal.

The average spectrum is a spectrum of the relevant part or parts of the speech obtained and averaged over multiple frames.

10

- The spectrum or spectra can be averaged over enough data to provide reasonable confidence in the information average. In general terms, this average will become more reliable as more data is used to form the average spectrum or spectra. In some cases, spectra averaged over 500ms of the relevant speech will be enough to provide reliable averaged spectra. The length of time over which the averaged spectrum or spectra are generated may be adapted based on the articulation rate of the speech, in order to ensure that the speech contains enough phonetic variation to provide a reliable average. The length of time over which the averaged spectrum or spectra are generated may be adapted based on the content of the speech.
- 15

20

As mentioned above, an average spectrum of at least a part of the cleaned audio signal is obtained in step 164. For example, this may comprise obtaining an average spectrum for parts of the cleaned audio signal representing one or more audio classes.

- 25 To achieve this, one or more components of the cleaned audio signal, representing different acoustic classes of the speech, are extracted from the cleaned audio signal. Extracting the or each component of the cleaned audio signal may comprise identifying periods when the cleaned audio signal contains the relevant acoustic class of speech. More specifically, extracting the component or components of the cleaned audio signal may comprise identifying frames of the cleaned audio signal that contain the relevant acoustic class of speech.
- 30

- In some embodiments, obtaining an average spectrum of at least a part of the cleaned audio signal comprises obtaining an average spectrum of a part of the cleaned audio signal representing voiced speech.
- 35

In some other embodiments, obtaining an average spectrum of at least a part of the cleaned audio signal comprises obtaining a first average spectrum of a part of the cleaned audio signal representing voiced speech and obtaining a second average spectrum of a part of the cleaned audio signal representing unvoiced speech.

5

When the method involves obtaining an average spectrum for parts of the cleaned audio signal representing one or more audio classes, and the acoustic class is voiced speech (or the first and second acoustic classes of the speech are voiced speech and unvoiced speech), there are several methods that can be used to identify voiced and unvoiced speech, for example: using a deep neural network (DNN), trained against a golden reference, for example using Praat software; performing an autocorrelation with unit delay on the speech signal (because voiced speech has a higher autocorrelation for non-zero lags); performing a linear predictive coding (LPC) analysis (because the initial reflection coefficient is a good indicator of voiced speech); looking at the zero-crossing rate of the speech signal (because unvoiced speech has a higher zero-crossing rate); looking at the short term energy of the signal (which tends to be higher for voiced speech); tracking the first formant frequency F0 (because unvoiced speech does not contain the first format frequency); examining the error in a linear predictive coding (LPC) analysis (because the LPC prediction error is lower for voiced speech); using automatic speech recognition to identify the words being spoken and hence the division of the speech into voiced and unvoiced speech; or fusing any or all of the above.

As mentioned above, the acoustic classes of the speech may be voiced speech and unvoiced speech. However, the acoustic classes of the speech may be any phonetically distinguishable acoustic classes. For example, they may be different phoneme classes, for example two different sets of vowels; they may be two different fricatives; or a first class may be fricatives while a second class are sibilants.

In step 166 of the method, the obtained average spectrum of at least a part of the cleaned audio signal is passed to a comparison block 184. The comparison block 184 also receives one or more long term average speaker model for one or more enrolled speaker. The term "long term" average speaker model means that enough of the speech of the enrolled speaker was used to form the model, either during enrolment or subsequently, that the model is relatively stable.

In some embodiments or situations, there is only one enrolled speaker, and so the comparison block 184 receives the one or more long term average speaker model for that enrolled speaker. In some other embodiments or situations, there is more than one enrolled speaker, and so the comparison block 184 receives the one or more long term average speaker model for each enrolled speaker.

In some other embodiments or situations, there is more than one enrolled speaker, but there is some additional information regarding the purported speaker. For example, a user of the device may have identified themselves in some way. In that case, the comparison block 184 receives the one or more long term average speaker model for that enrolled speaker.

In addition, in some embodiments, the comparison block 184 may additionally or alternatively receive a Universal Background Model (UBM), for example in the form of a model of the statistically average user.

The one or more long term average speaker model, and the Universal Background Model (UBM) if used, are stored in a model database 186.

The comparison block 184 may receive one or more long term average speaker model corresponding to the part of the cleaned audio signal for which the average spectrum was obtained.

Thus, for example, obtaining an average spectrum of at least a part of the cleaned audio signal may comprise obtaining an average spectrum of a part of the cleaned audio signal representing voiced speech. That is, with a measurement S_v of the spectrum of the voiced speech, and with values having been calculated for the channel, α , and for the noise, n , the cleaned measurement SC_v of the spectrum of the voiced speech can be calculated as:

$$SC_v = \frac{(S_v - n)}{\alpha}.$$

This can then be compared with the long term average speaker model M_v for voiced speech of the or each enrolled speaker being considered by the comparison block 184.

In other examples, obtaining an average spectrum of at least a part of the cleaned audio signal may comprise obtaining a first average spectrum of a part of the cleaned

audio signal representing voiced speech and obtaining a second average spectrum of a part of the cleaned audio signal representing unvoiced speech.

As before, the average spectrum of a part of the cleaned audio signal representing
 5 voiced speech can be calculated as:

$$SC_v = \frac{(S_v - n)}{\alpha},$$

and similarly the average spectrum of a part of the cleaned audio signal representing unvoiced speech can be calculated as:

$$SC_u = \frac{(S_u - n)}{\alpha}.$$

10

The first average spectrum SC_v is compared with a long term average speaker model M_v for voiced speech of the or each enrolled speaker being considered by the comparison block 184, and the second average spectrum SC_u is compared with a long term average speaker model M_u for unvoiced speech of the or each enrolled
 15 speaker being considered by the comparison block 184.

In step 168 of the method, the result of the comparison is passed to a determination block 188, which determines based on the comparison whether the speech is the speech of the enrolled speaker being considered by the comparison block 184. As
 20 mentioned above, this determination may be an accept/reject decision based on the comparison, as to whether the received speech matches sufficiently closely with the enrolled user who was expected to be the speaker.

In some examples, a small number of speakers (for example from 2 to 10) are enrolled,
 25 and suitable models of their speech are obtained during an enrolment process. Then, the determination made by the determination block 188 concerns which of those enrolled speakers was the most likely candidate as the source of the speech in the received audio signal.

30 This determination may be based on the respective Log Spectral Distances (LSD) of the received speech from the different models, or may use Principal component analysis (PCA) or Linear discriminative analysis (LDA), as examples.

When a Universal Background Model (UBM) is also considered, then the determination may take into account the result of a comparison between the received speech, the model of the enrolled user's speech, and the background model.

5 **Figure 14** is another block diagram of a system for performing speaker identification.

As described above, the system may be implemented in a smartphone, such as the smartphone 10, or any other device with voice biometric functionality. Thus, the smartphone 10 is able to perform various functions in response to spoken commands
10 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
15 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
20 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
25 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.

30 Some embodiments are particularly suited to use in devices, such as home control systems, home entertainment systems, or in-vehicle entertainment systems, in which there will often be multiple enrolled users (for example between two and ten such users), and where the intended operation to be performed in response to a spoken command (such as "play my favourite music", or "increase the temperature in my
35 room", for example) will depend on the identity of the speaker.

As in the system of Figure 13, the signal generated by a microphone 12 in response to ambient sound is received. The received signal is divided into frames, which may for example have lengths in the range of 10-100 ms. These frames can be analysed to determine whether they represent speech, and only frames that represent speech are
5 considered further.

Components of the received audio signal, representing different acoustic classes of the speech, are then extracted in an extraction block 192. Extracting the or each component of the cleaned audio signal may comprise identifying periods when the
10 audio signal contains the relevant acoustic class of speech. More specifically, extracting the component or components of the audio signal may comprise identifying frames of the audio signal that contain the relevant acoustic class of speech.

In the illustrated embodiment, the extraction block 192 is a voiced/unvoiced detector
15 (VU), which extracts respective components representing voiced and unvoiced speech, and outputs an average spectrum S_v of a part of the audio signal representing voiced speech, and an average spectrum S_u of a part of the audio signal representing unvoiced speech.

20 When the first and second acoustic classes of the speech are voiced speech and unvoiced speech, there are several methods that can be used to identify voiced and unvoiced speech, for example: using a deep neural network (DNN), trained against a golden reference, for example using Praat software; performing an autocorrelation with unit delay on the speech signal (because voiced speech has a higher autocorrelation
25 for non-zero lags); performing a linear predictive coding (LPC) analysis (because the initial reflection coefficient is a good indicator of voiced speech); looking at the zero-crossing rate of the speech signal (because unvoiced speech has a higher zero-crossing rate); looking at the short term energy of the signal (which tends to be higher for voiced speech); tracking the first formant frequency F_0 (because unvoiced speech
30 does not contain the first format frequency); examining the error in a linear predictive coding (LPC) analysis (because the LPC prediction error is lower for voiced speech); using automatic speech recognition to identify the words being spoken and hence the division of the speech into voiced and unvoiced speech; or fusing any or all of the above.

As mentioned above, the acoustic classes of the speech may be voiced speech and unvoiced speech. However, the acoustic classes of the speech may be any phonetically distinguishable acoustic classes. For example, they may be different phoneme classes, for example two different sets of vowels; they may be two different
 5 fricatives; or a first class may be fricatives while a second class are sibilants.

The average spectra of the two components of the signal representing the two acoustic classes of the speech are then passed to a channel/noise calculation and removal
 10 block 194.

In some embodiments, the system is provided with a purported identity of the speaker, and it is required to determine whether the received signal has in fact come from that speaker (referred to as speaker verification). In other embodiments, the system has multiple enrolled speakers, but has no further information as to which of the enrolled
 15 speakers is speaking at any given time, and it is required to identify which of those enrolled speakers is the speaker (referred to as speaker identification).

The system includes a database 196, which stores a long term average speaker model M_v for voiced speech of the or each enrolled speaker and a long term average
 20 speaker model M_u for unvoiced speech of the or each enrolled speaker (or models of other acoustic classes of the speech of each enrolled speaker).

As described above, the system may be required to perform speaker verification, or speaker identification.
 25

In the case of speaker verification, the average spectrum S_v of the part of the audio signal representing voiced speech, and the average spectrum S_u of the part of the audio signal representing unvoiced speech, are combined with the model M_v for
 30 M_u for unvoiced speech of the purported speaker to obtain values for the channel, α , and for the noise, n . Specifically, as before:

$$\alpha = \frac{(S_u - S_v)}{(M_u - M_v)}, \text{ and}$$

$$n = \frac{(SuMv - SvMu)}{(Mu - Mv)}.$$

The channel/noise calculation and removal block 194 then removes the effect of the
 5 calculated channel and noise, to obtain a cleaned measurement SCv of the average
 spectrum of the voiced speech, calculated as:

$$SCv = \frac{(Sv - n)}{\alpha}.$$

In other embodiments, a cleaned measurement SCu of the average spectrum of the
 10 unvoiced speech can be similarly calculated as:

$$SCu = \frac{(Su - n)}{\alpha}.$$

The cleaned measurement of the average spectrum of the relevant part of the speech
 is then passed to a comparison block 198, for comparison with the respective model of
 15 that part of the speech of the purported user. The comparison score is output,
 indicating whether the cleaned measurement(s) of the average spectrum of the
 relevant part(s) of the speech is/are close enough to the model(s) to have a required
 degree of confidence that the signal comes from the speech of the purported speaker.
 As before, the comparison block 198 may additionally receive a Universal Background
 20 Model (UBM), for example in the form of a model of the statistically average user, from
 the database 196, and may use this when providing the output comparison score.

In the case of speaker identification, the average spectrum Sv of the part of the audio
 25 signal representing voiced speech, and the average spectrum Su of the part of the
 audio signal representing unvoiced speech, are combined with the respective models
 Mv for voiced speech of each enrolled speaker and the long term average speaker
 model Mu for unvoiced speech of each enrolled speaker to obtain preliminary or
 hypothetical values for the channel, α , and for the noise, n . Specifically, as before:

30

$$\alpha = \frac{(Su - Sv)}{(Mu - Mv)}, \text{ and}$$

$$n = \frac{(SuMv - SvMu)}{(Mu - Mv)}.$$

These values for channel and noise are calculated for each of the possible speakers.

- 5 The results may be such that it is clear that the speech could not have come from one or more of the enrolled speakers. Specifically, if the calculated values for the channel, α , based on the models for a particular speaker, are clearly physically implausible, it can be assumed that that speaker was not the source of the received speech signal. For example, if there are very large variations (of more than 20dB, say) in one of the
- 10 calculated channels across the relevant frequency range, or if there are significant discontinuities in one of the calculated channels, this might indicate that that channel is physically implausible, and hence that the speaker whose model led to that calculated channel was not the person speaking at that time.

- 15 Otherwise, the channel/noise calculation and removal block 194 removes the effect of each of the calculated channel and noise values from the received signal, to obtain respective cleaned hypothetical measurements SCv of the average spectrum of the voiced speech, on the assumption that the speaker was the person whose speech model was used as the basis for those calculated values of the channel and noise.

20

Thus, in a case with two enrolled speakers A and B , having respective models MvA and MvB for their voiced speech and having respective models MuA and MuB for their unvoiced speech, it is possible to obtain respective hypothetical values for the channel and noise, namely:

25

$$\alpha A = \frac{(Su - Sv)}{(MuA - MvA)}, \text{ and}$$

$$nA = \frac{(SuMvA - SvMuA)}{(MuA - MvA)} \text{ for enrolled speaker } A, \text{ and}$$

$$\alpha B = \frac{(Su - Sv)}{(MuA - MvB)}, \text{ and}$$

$$nB = \frac{(SuMvB - SvMuB)}{(MuB - MvB)} \text{ for enrolled speaker } B.$$

30

These are then provisionally removed from the received signal to give respective hypothetical cleaned measurements for the two enrolled users, namely:

$$SCvA = \frac{(Sv - nA)}{\alpha A} \text{ for enrolled speaker } A \text{ and}$$

5

$$SCvB = \frac{(Sv - nB)}{\alpha B} \text{ for enrolled speaker } B.$$

These hypothetical cleaned measurements of the average spectrum of the relevant part of the speech are then passed to a comparison block 198, for comparison with the
10 respective model of that part of the speech of the relevant user.

Thus, $SCvA$ is compared with the model MvA , for enrolled speaker A , and $SCvB$ is compared with the model MvB for enrolled speaker B .

15 The comparison score is then output, indicating whether the hypothetical cleaned measurement of the average spectrum of the relevant part of the speech for one of the enrolled speakers is close enough to the respective model to have a required degree of confidence that the signal comes from the speech of that enrolled speaker.

20 The result output by the comparison block 198 may simply indicate which of those enrolled speakers was the most likely candidate as the source of the speech in the received audio signal.

The skilled person will recognise that some aspects of the above-described apparatus
25 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
30 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-
5 (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
10 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
15 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
20 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
25 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
30 numerals or labels in the claims shall not be construed so as to limit their scope.

CLAIMS

1. A method of analysis of an audio signal, the method comprising:
receiving an audio signal representing speech;
5 extracting first and second components of the audio signal representing first and second acoustic classes of the speech respectively;
analysing the first and second components of the audio signal with models of the first and second acoustic classes of the speech of an enrolled user; and
based on said analysing, obtaining information about at least one of a channel
10 and noise affecting said audio signal.
2. A method according to claim 1, wherein extracting first and second components of the audio signal comprises:
identifying periods when the audio signal contains voiced speech; and
15 identifying remaining periods of speech as containing unvoiced speech.
3. A method according to claim 1 or 2, wherein analysing the first and second components of the audio signal with the models of the first and second acoustic classes of the speech of the enrolled user comprises:
20 comparing magnitudes of the audio signal at a number of predetermined frequencies with magnitudes in the models of the first and second acoustic classes of the speech.
4. A method according to any preceding claim, comprising compensating the
25 received audio signal for channel and/or noise.
5. A method according to any preceding claim, comprising:
performing a speaker identification process on the received audio signal to form a provisional decision on an identity of a speaker;
30 selecting the models of the first and second acoustic classes of the speech of the enrolled user, from a plurality of models, based on the provisional decision on the identity of the speaker;
compensating the received audio signal for channel and/or noise; and
performing a second speaker identification process on the compensated received
35 audio signal to form a final on the identity of the speaker.

6. A method according to claim 5, wherein compensating the received audio signal for channel and/or noise comprises:
- identifying at least one part of a frequency spectrum of the received audio signal where a noise level exceeds a threshold level; and
- 5 ignoring the identified part of the frequency spectrum of the received audio signal when performing the second speaker identification process.
7. A method according to any of claims 1 to 6, wherein the first and second acoustic classes of the speech comprise voiced speech and unvoiced speech.
- 10
8. A method according to any of claims 1 to 6, wherein the first and second acoustic classes of the speech comprise first and second phoneme classes.
9. A method according to any of claims 1 to 6, wherein the first and second acoustic
- 15 classes of the speech comprise first and second fricatives.
10. A method according to any of claims 1 to 6, wherein the first and second acoustic classes of the speech comprise fricatives and sibilants.
- 20
11. A system for analysis of an audio signal, the system comprising an input for receiving an audio signal, and being configured for:
- receiving an audio signal representing speech;
- extracting first and second components of the audio signal representing first and second acoustic classes of the speech respectively;
- 25 analysing the first and second components of the audio signal with models of the first and second acoustic classes of the speech of an enrolled user; and
- based on said analysing, obtaining information about at least one of a channel and noise affecting said audio signal.
- 30
12. A device comprising a system as claimed in any of claims 1 to 10.
13. A device as claimed in claim 12, 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
- 35 or a domestic appliance.

14. A computer program product, comprising a computer-readable tangible medium, and instructions for performing a method according to any one of claims 1 to 10.
15. 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 10.
16. A method of determining whether a received signal may result from a replay attack, the method comprising:
- 10 receiving an audio signal representing speech;
obtaining information about a channel affecting said audio signal; and
determining whether the channel has at least one characteristic of a loudspeaker.
17. A method according to claim 16, wherein determining whether the channel has at least one characteristic of a loudspeaker comprises:
- 15 determining whether the channel has a low frequency roll-off.
18. A method according to claim 17, wherein determining whether the channel has a low frequency roll-off comprises determining whether the channel decreases at a constant rate for frequencies below a lower cut-off frequency.
- 20
19. A method according to claim 16 or 17, wherein determining whether the channel has at least one characteristic of a loudspeaker comprises:
- determining whether the channel has a high frequency roll-off.
- 25
20. A method according to claim 19, wherein determining whether the channel has a high frequency roll-off comprises determining whether the channel decreases at a constant rate for frequencies above an upper cut-off frequency.
- 30
21. A method according to claim 16, 17 or 19, wherein determining whether the channel has at least one characteristic of a loudspeaker comprises:
- determining whether the channel has ripple in a pass-band thereof.
22. A method according to claim 21, wherein determining whether the channel has ripple in a pass-band thereof comprises determining whether a degree of ripple over a
- 35

central part of the pass-band, for example from 100Hz – 10kHz, exceeds a threshold amount.

23. A system for determining whether a received signal may result from a replay
5 attack, the system comprising an input for receiving an audio signal, and being
configured for:
- receiving an audio signal representing speech;
 - obtaining information about a channel affecting said audio signal; and
 - 10 determining whether the channel has at least one characteristic of a loudspeaker.

24. A device comprising a system as claimed in any of claims 16 to 22.

25. A device as claimed in claim 24, wherein the device comprises a mobile
telephone, an audio player, a video player, a mobile computing platform, a games
15 device, a remote controller device, a toy, a machine, or a home automation controller
or a domestic appliance.

26. A computer program product, comprising a computer-readable tangible medium,
and instructions for performing a method according to any one of claims 16 to 22.

27. 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 16 to 22.

28. A method of speaker identification, comprising:
- 25 receiving an audio signal representing speech;
 - removing effects of a channel and/or noise from the received audio signal to
obtain a cleaned audio signal;
 - obtaining an average spectrum of at least a part of the cleaned audio signal;
 - comparing the average spectrum with a long term average speaker model for an
30 enrolled speaker; and
 - determining based on the comparison whether the speech is the speech of the
enrolled speaker.

29. A method according to claim 28, wherein obtaining an average spectrum of at
35 least a part of the cleaned audio signal comprises obtaining an average spectrum of a
part of the cleaned audio signal representing voiced speech.

30. A method according to claim 28, wherein obtaining an average spectrum of at least a part of the cleaned audio signal comprises obtaining a first average spectrum of a part of the cleaned audio signal representing a first acoustic class and obtaining a
5 second average spectrum of a part of the cleaned audio signal representing a second acoustic class, and wherein
comparing the average spectrum with a long term average speaker model for an enrolled speaker comprises comparing the first average spectrum with a long term average speaker model for the first acoustic class for the enrolled speaker and
10 comparing the second average spectrum with a long term average speaker model for the second acoustic class for the enrolled speaker.
31. A method according to claim 28, wherein the first acoustic class is voiced speech and the second acoustic class is unvoiced speech.
15
32. A method according to claim 28, 29, 30 or 31, comprising comparing the average spectrum with respective long term average speaker models for each of a plurality of enrolled speakers; and
determining based on the comparison whether the speech is the speech of one
20 of the enrolled speakers.
33. A method according to claim 32, further comprising comparing the average spectrum with a Universal Background Model; and
including a result of the comparing the average spectrum with the Universal
25 Background Model in determining whether the speech is the speech of one of the enrolled speakers.
34. A method according to claim 32, comprising identifying one of the enrolled speakers as a most likely candidate as a source of the speech.
30
35. A method according to any of claims 28 to 34, comprising:
obtaining information about the effects of a channel and/or noise on the received audio signal by:
receiving the audio signal representing speech;
35 extracting first and second components of the audio signal representing first and second acoustic classes of the speech respectively;

analysing the first and second components of the audio signal with models of the first and second acoustic classes of the speech of an enrolled user; and

based on said analysing, obtaining information about at least one of a channel and noise affecting said audio signal.

5

36. A method according to claim 35, comprising analysing the first and second components of the audio signal with models of the first and second acoustic classes of the speech of a plurality of enrolled users, to obtain respective hypothetical values of the channel, and determining that the speech is not the speech of any enrolled speaker
10 whose models give rise to physically implausible hypothetical values of the channel.

15

37. A method according to claim 36, wherein a hypothetical value of the channel is considered to be physically implausible if it contains variations exceeding a threshold level across the relevant frequency range.

38. A method according to claim 36, wherein a hypothetical value of the channel is considered to be physically implausible if it contains significant discontinuities.

39. A system for analysis of an audio signal, the system comprising an input for
20 receiving an audio signal, and being configured for:
receiving an audio signal representing speech;
removing effects of a channel and/or noise from the received audio signal to obtain a cleaned audio signal;
obtaining an average spectrum of at least a part of the cleaned audio signal;
25 comparing the average spectrum with a long term average speaker model for an enrolled speaker; and
determining based on the comparison whether the speech is the speech of the enrolled speaker.

30 40. A device comprising a system as claimed in claim 39.

41. A device as claimed in claim 40, 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
35 or a domestic appliance.

42. A computer program product, comprising a computer-readable tangible medium, and instructions for performing a method according to any one of claims 28 to 38.

43. A non-transitory computer readable storage medium having computer-executable
5 instructions stored thereon that, when executed by processor circuitry, cause the processor circuitry to perform a method according to any one of claims 28 to 38.

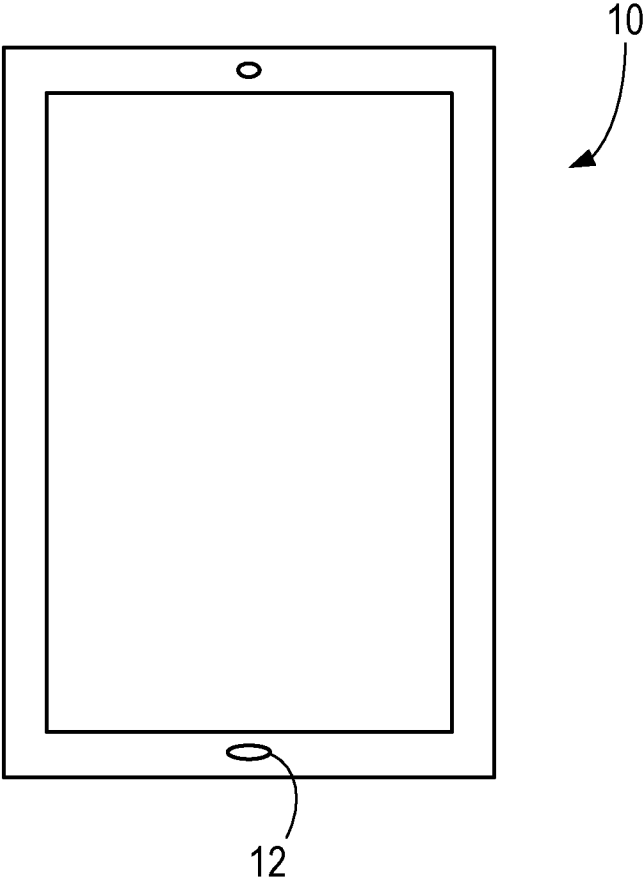


Figure 1

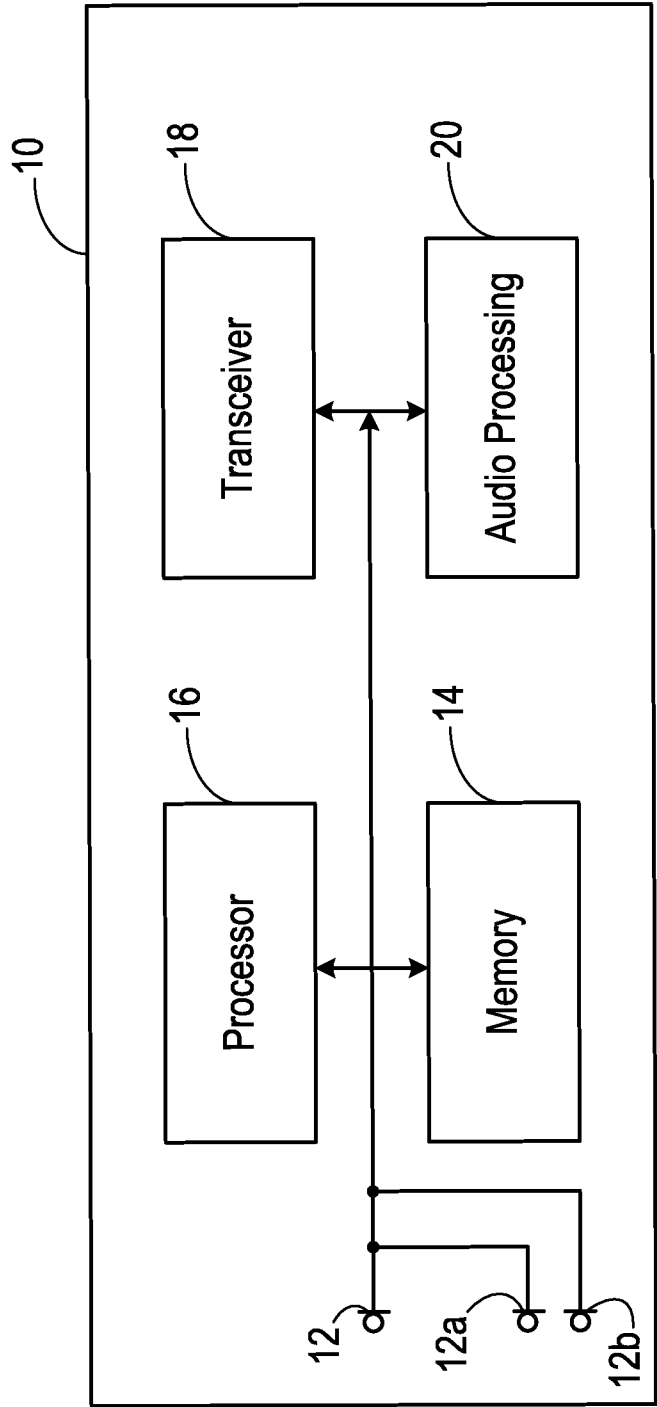


Figure 2

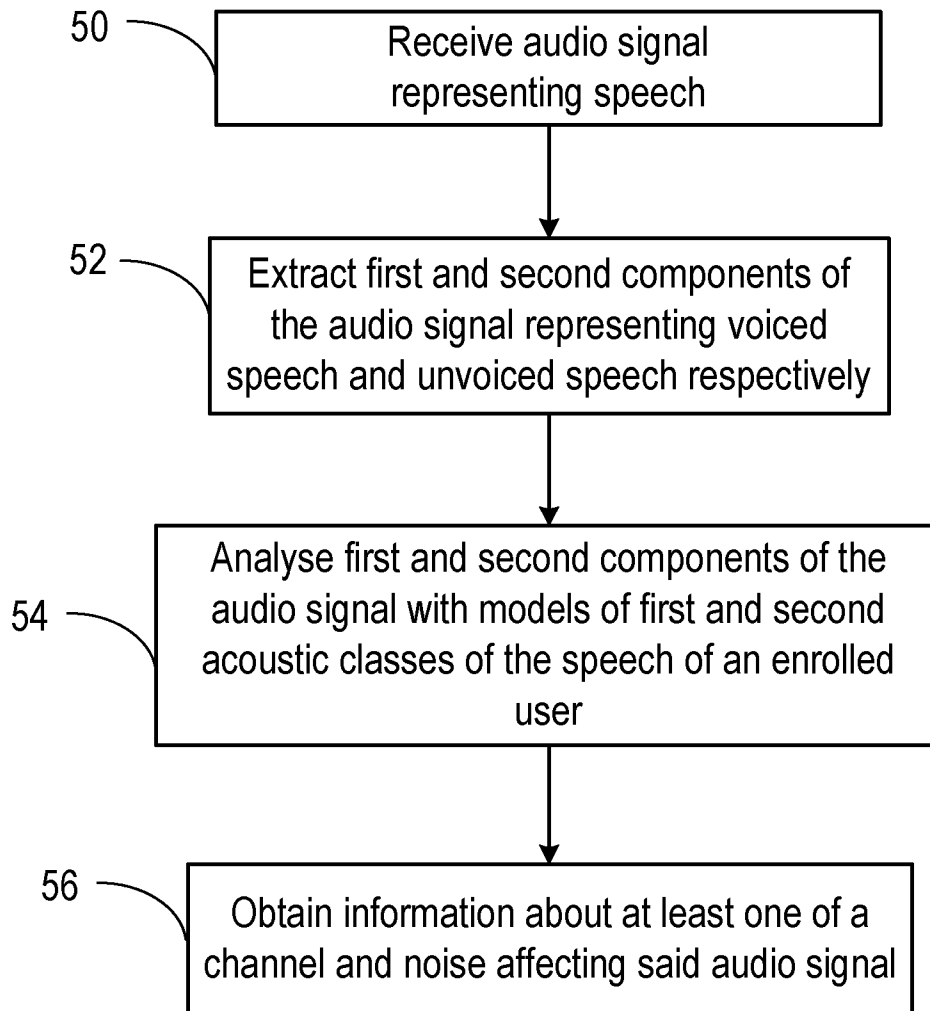


Figure 3

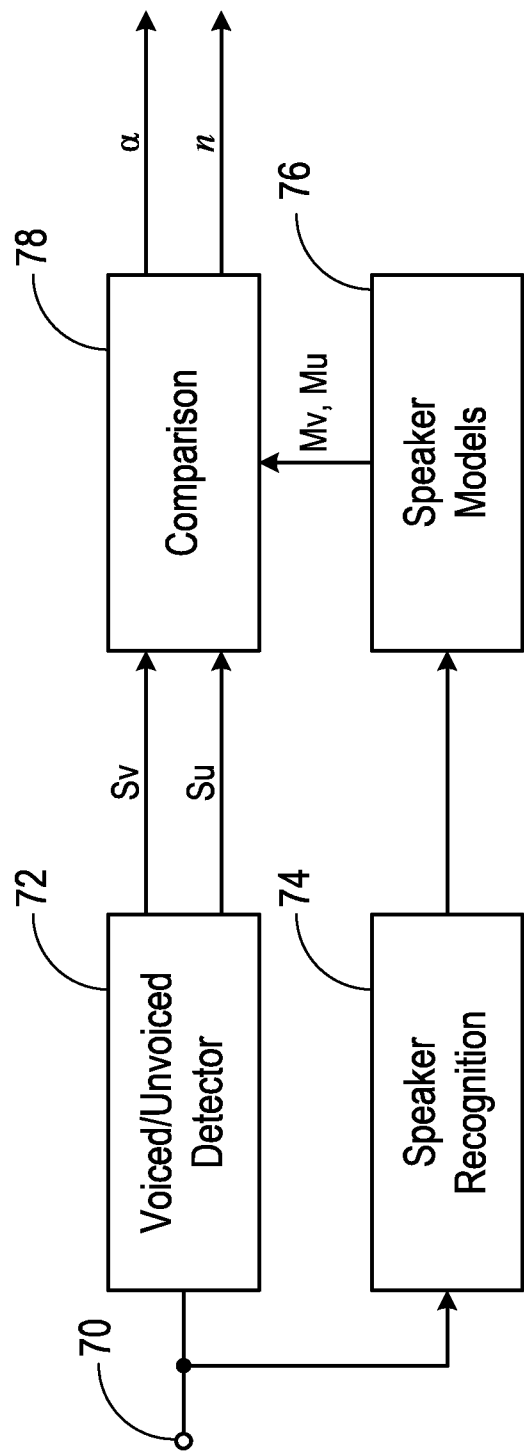


Figure 4

5/14

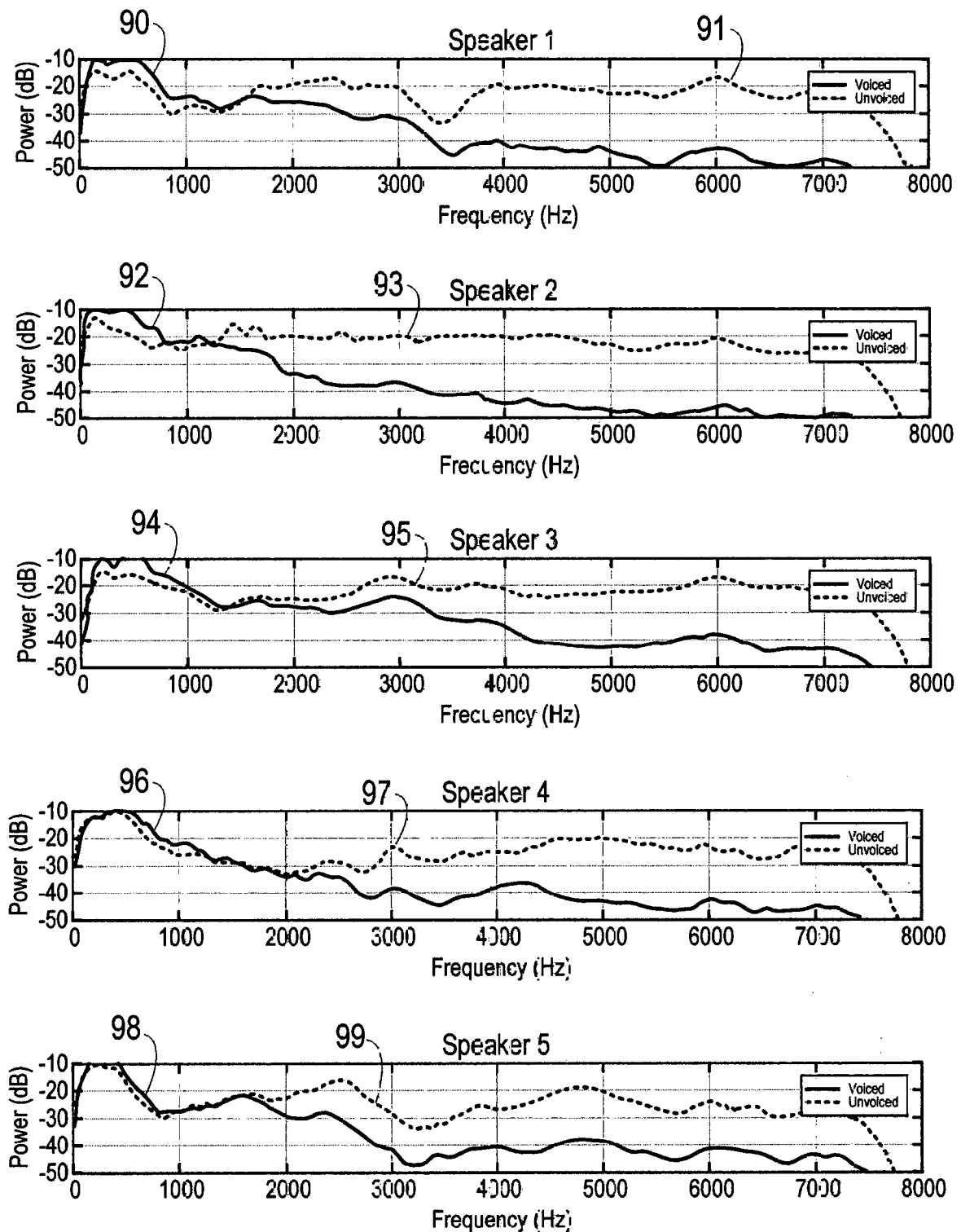


Figure 5

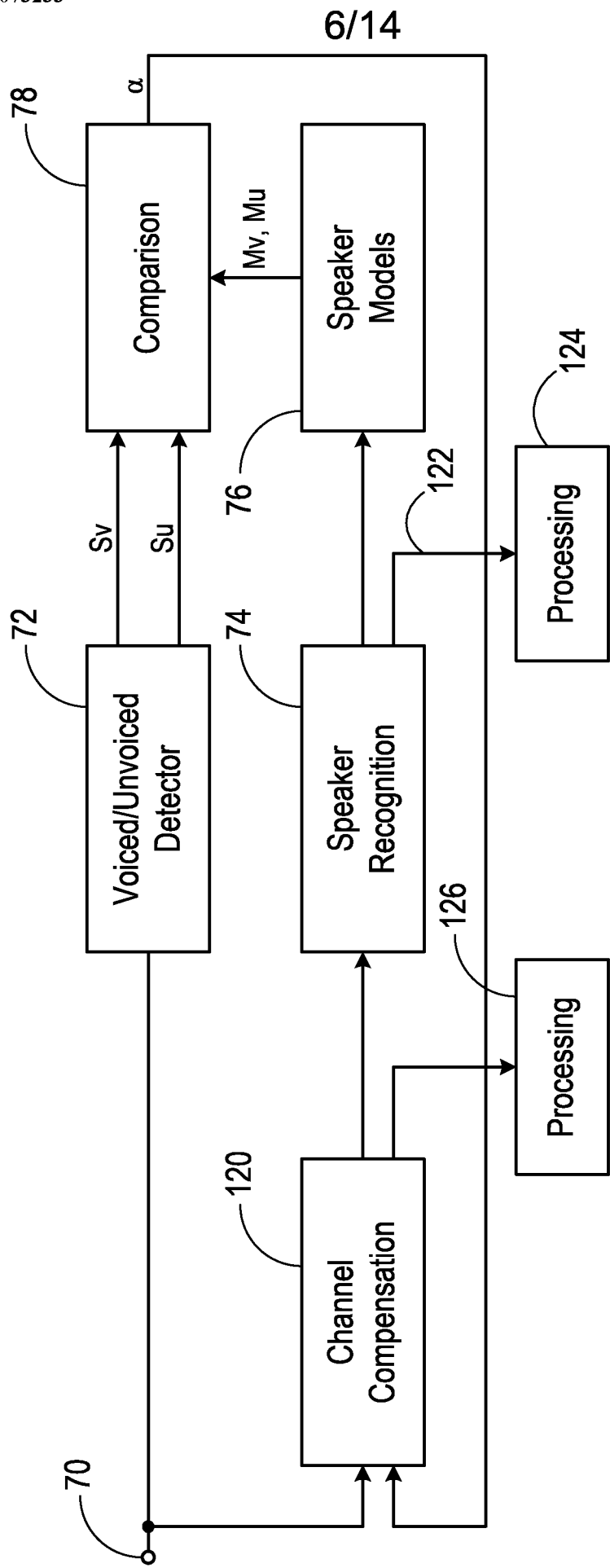


Figure 6

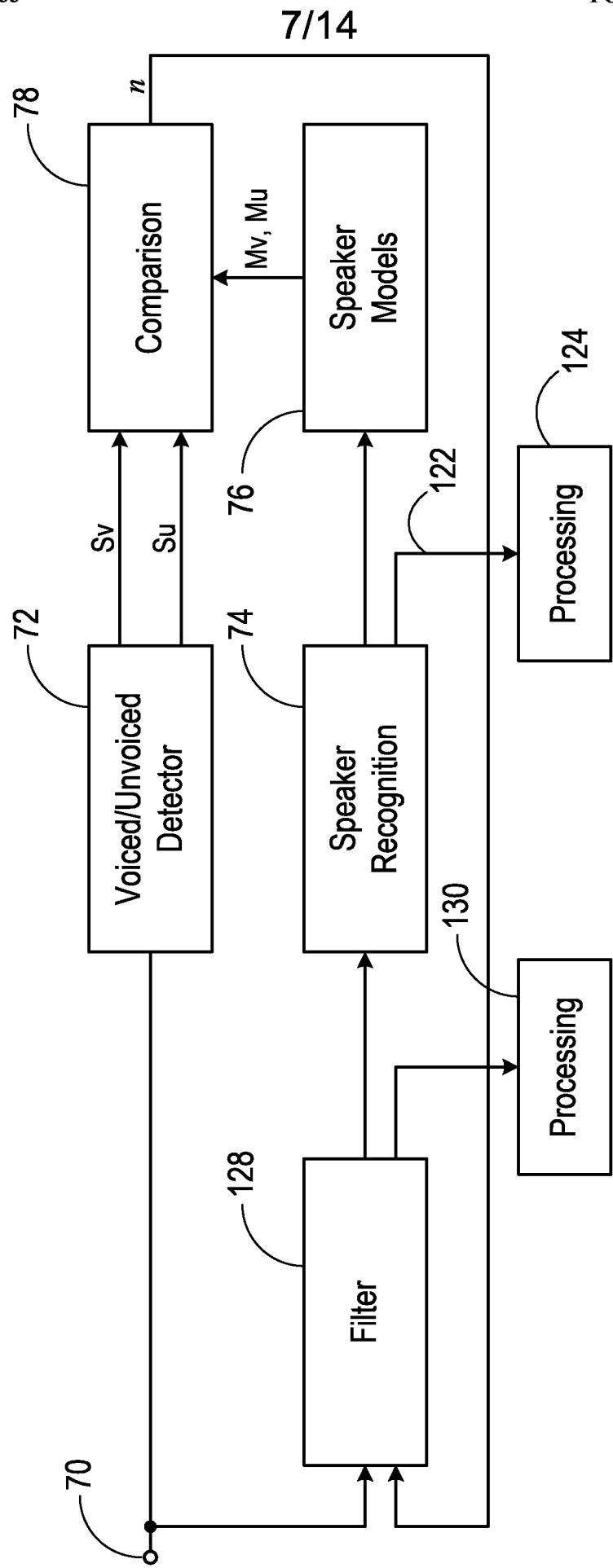


Figure 7

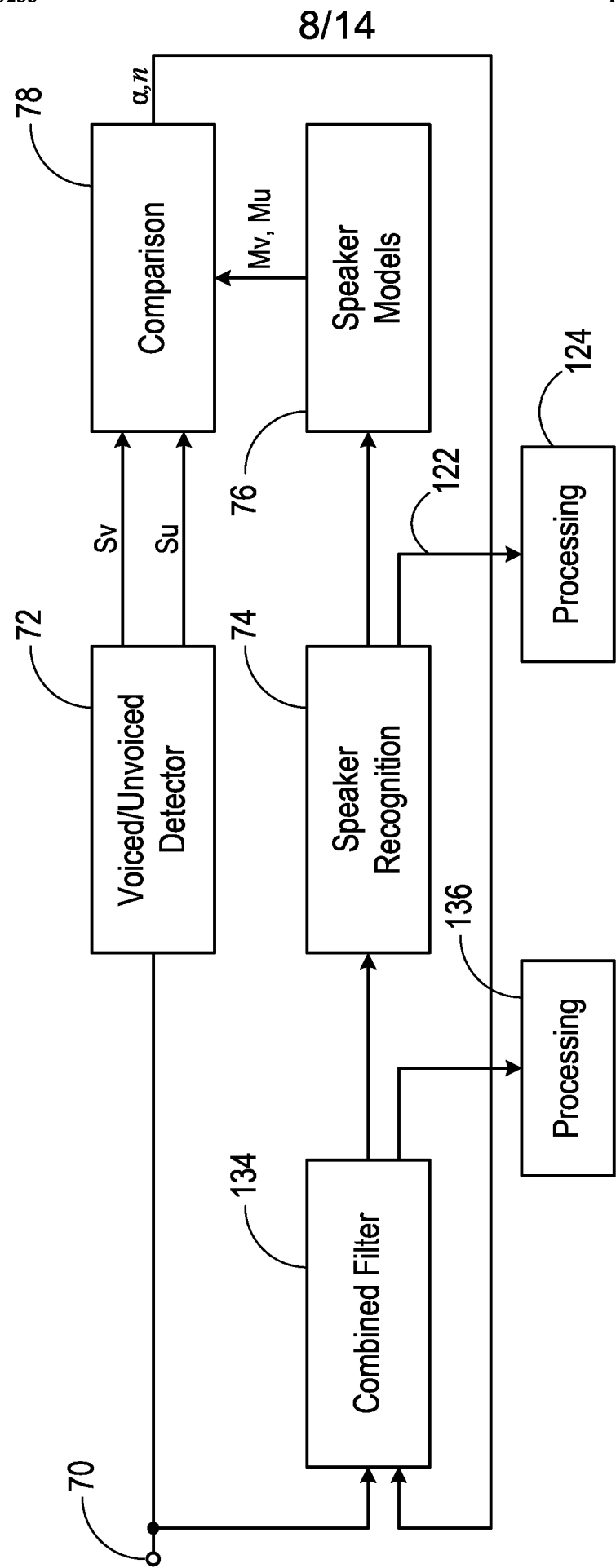


Figure 8

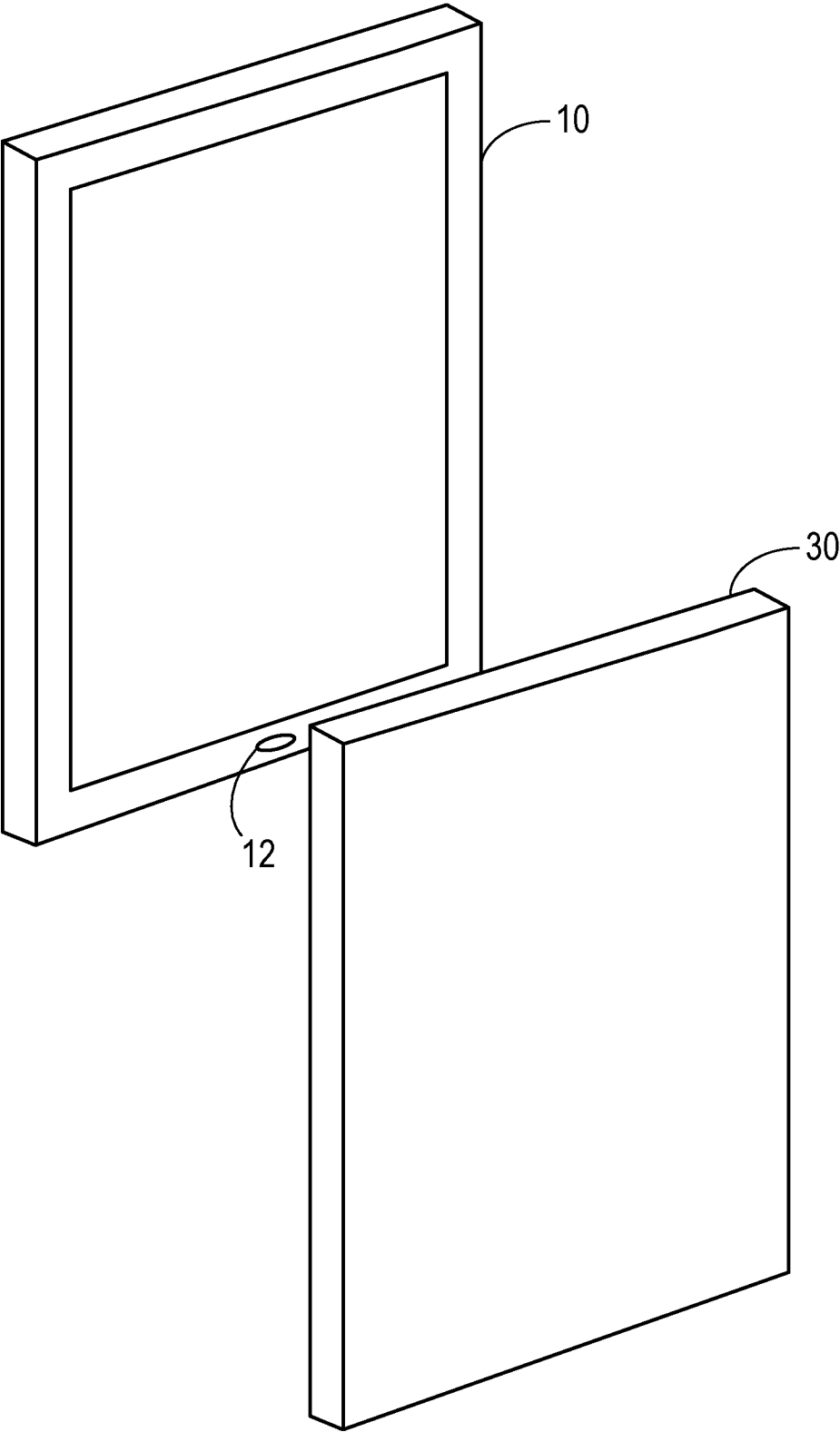


Figure 9

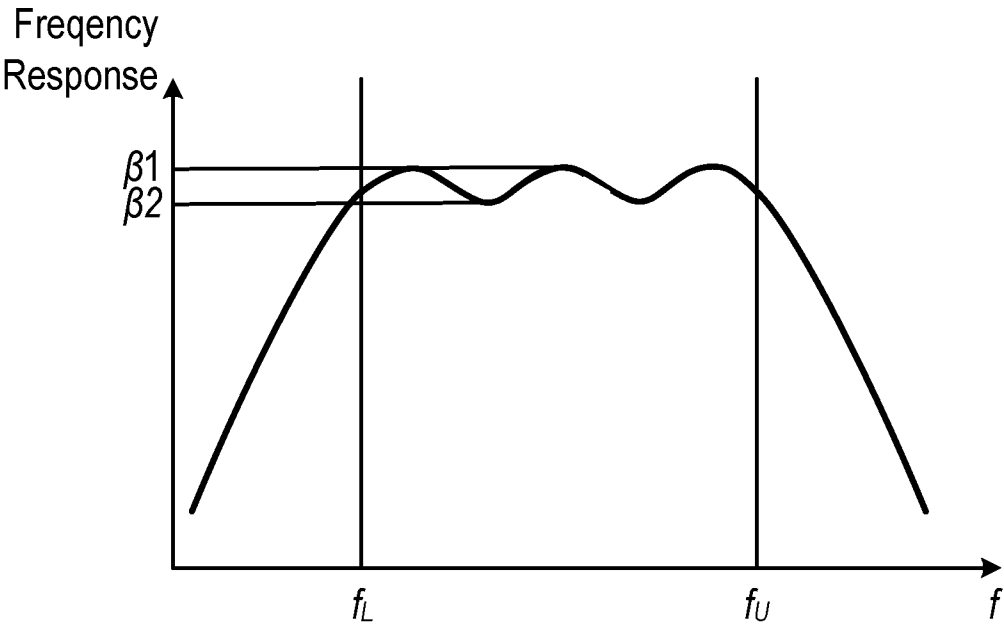


Figure 10

11/14

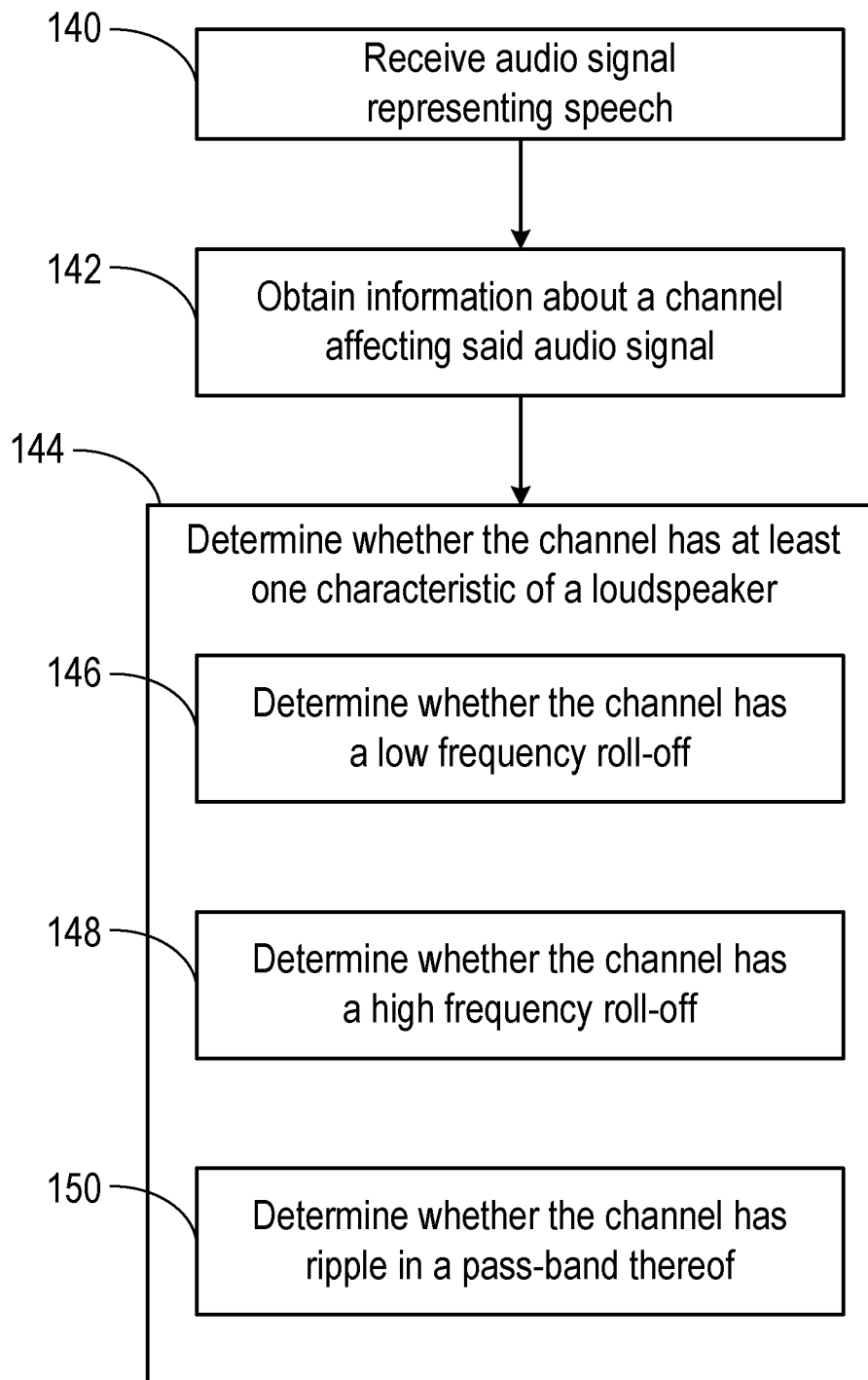


Figure 11

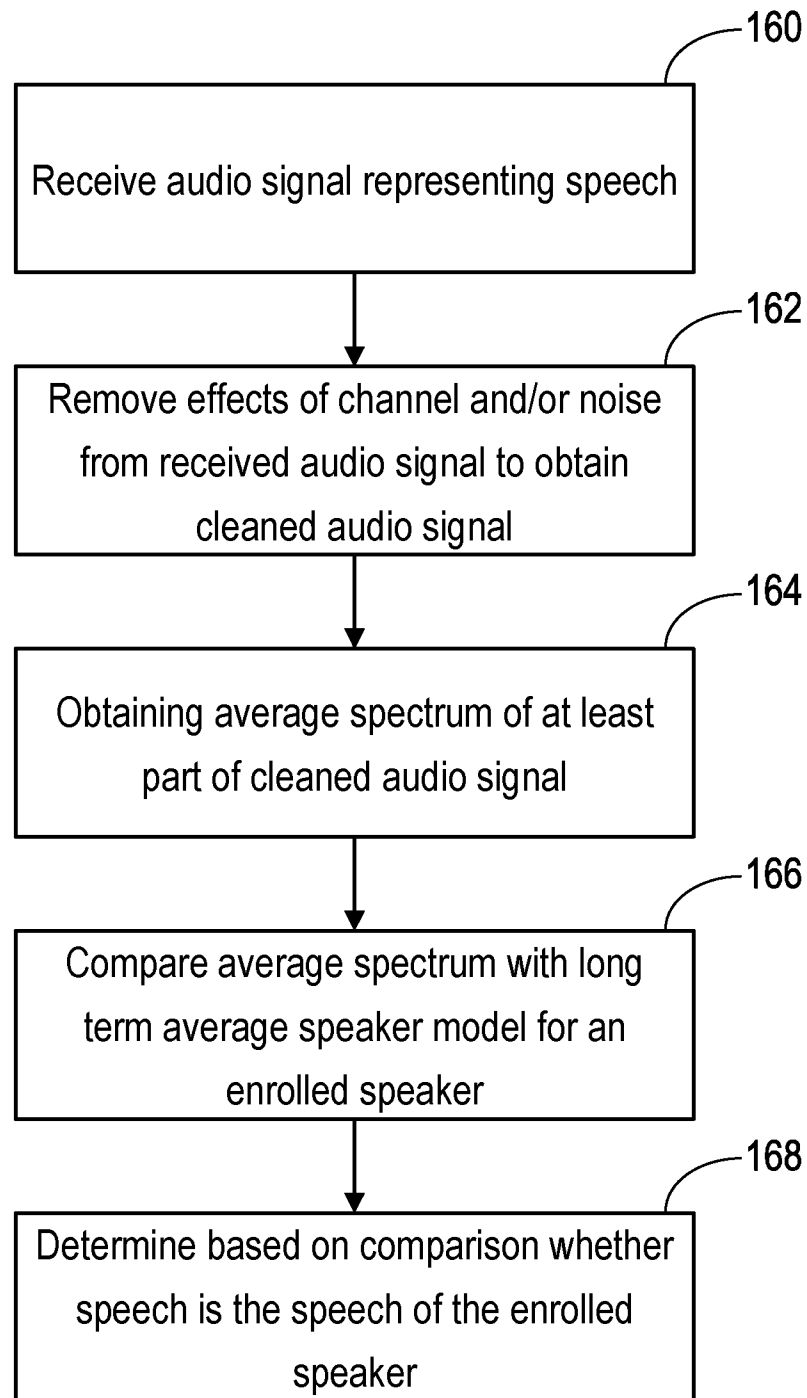


Figure 12

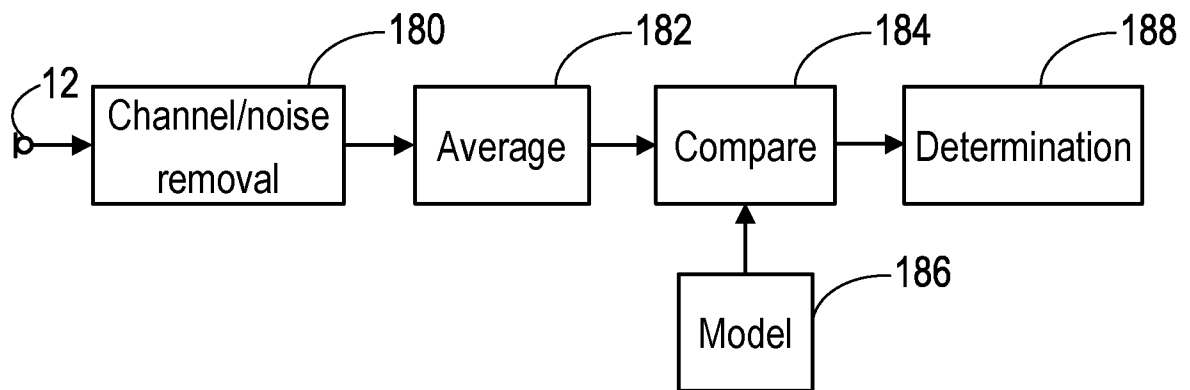


Figure 13

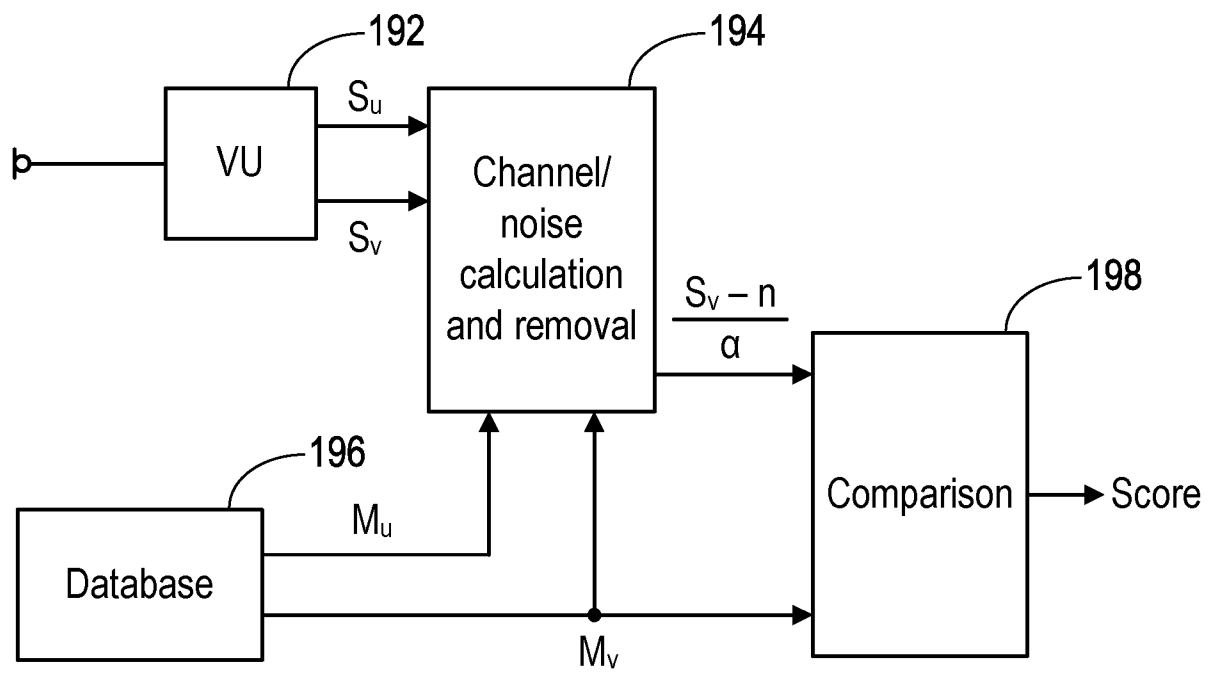


Figure 14

INTERNATIONAL SEARCH REPORT

International application No

PCT/GB2018/052905

A. CLASSIFICATION OF SUBJECT MATTER

INV. G10L25/93 G10L17/14
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

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)

EP0-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2007/129941 A1 (TAVARES CLIFFORD [US]) 7 June 2007 (2007-06-07) abstract paragraph [[0039]] paragraphs [[0040]], [[0059]] paragraphs [[0071]] - [[0075]] figures 5,6,10 -----	1-15
X	EP 2 860 706 A2 (AGNITIO S L [ES]) 15 April 2015 (2015-04-15) cited in the application	16-23, 26,27
A	abstract paragraphs [[0084]] - [[0086]] paragraphs [[0137]], [[0158]] figure 1d ----- -/-	24,25

☒ Further documents are listed in the continuation of Box C.

☒ 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

17 January 2019

Date of mailing of the international search report

19/03/2019

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Greiser, Norbert

INTERNATIONAL SEARCH REPORT

International application No.
PCT/GB2018/052905

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:
2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

see additional sheet

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of additional fees.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

1-27

Remark on Protest

- ☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.
- ☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
- ☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

International application No

PCT/GB2018/052905

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 2013/022930 A1 (INTELLISIS CORP [US]; BRADLEY DAVID C [US]; GOLDIN DANIEL S [US]; HILT) 14 February 2013 (2013-02-14) abstract paragraphs [[88]] - [[91]] paragraphs [[105]] - [[107]] paragraphs [[113]] - [[114]] figure 1 -----	1-16,23, 24,26,27
A	WO 02/103680 A2 (SECURIVOX LTD [GB]; SAPELUK ANDREW THOMAS [GB]) 27 December 2002 (2002-12-27) abstract page 10, line 24 - page 11, line 21 figure 4 -----	1,11,12, 14,15

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/GB2018/052905

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2007129941 A1	07-06-2007	JP 2007156422 A US 2007129941 A1	21-06-2007 07-06-2007
EP 2860706 A2	15-04-2015	NONE	
WO 2013022930 A1	14-02-2013	CA 2847689 A1 CN 103999076 A EP 2742435 A1 HK 1199122 A1 HK 1199506 A1 KR 20140079369 A US 2013041658 A1 US 2014037095 A1 WO 2013022930 A1	14-02-2013 20-08-2014 18-06-2014 19-06-2015 03-07-2015 26-06-2014 14-02-2013 06-02-2014 14-02-2013
WO 02103680 A2	27-12-2002	AT 426234 T AU 2002311452 B2 CA 2451401 A1 CN 1543641 A EP 1399915 A2 US 2004236573 A1 WO 02103680 A2	15-04-2009 19-06-2008 27-12-2002 03-11-2004 24-03-2004 25-11-2004 27-12-2002

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

This International Searching Authority found multiple (groups of) inventions in this international application, as follows:

1. claims: 1-27

Method of analysis of an audio signal according to claim 1,
and method of determining whether a received signal may
result from a replay attack according to claim 16

1.1. claims: 1-15

Method of analysis of an audio signal

1.2. claims: 16-27

Method of determining whether a received signal may result
from a replay attack

2. claims: 28-43

Method of speaker identification
