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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
20 December 2018 (20.12.2018)

WIPO | PCT

(10) International Publication Number
WO 2018/229503 A1

(51) International Patent Classification:
H04R 1/10 (2006.01) *H04R 3/00* (2006.01)
H04R 1/46 (2006.01) *G10L 21/0216* (2013.01)
G10L 21/0208 (2013.01)

(21) International Application Number:
PCT/GB2018/051658

(22) International Filing Date:
15 June 2018 (15.06.2018)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/520,713 16 June 2017 (16.06.2017) US

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

(72) Inventors: **WATTS, David Leigh**; Level 1, Building 10, 658 Church Street, Cremorne, Victoria 3121 (AU). **STEELE, Brenton Robert**; Level 1, Building 10, 658 Church Street, Cremorne, Victoria 3121 (AU). **HARVEY, Thomas Ivan**; Level 1, Building 10, 658 Church Street, Cremorne, Victoria 3121 (AU). **SAPOZHNYKOV, Vitaliy**; Level 1, Building 10, 658 Church Street, Cremorne, Victoria 3121 (AU).

(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.

(54) Title: EARBUD SPEECH ESTIMATION

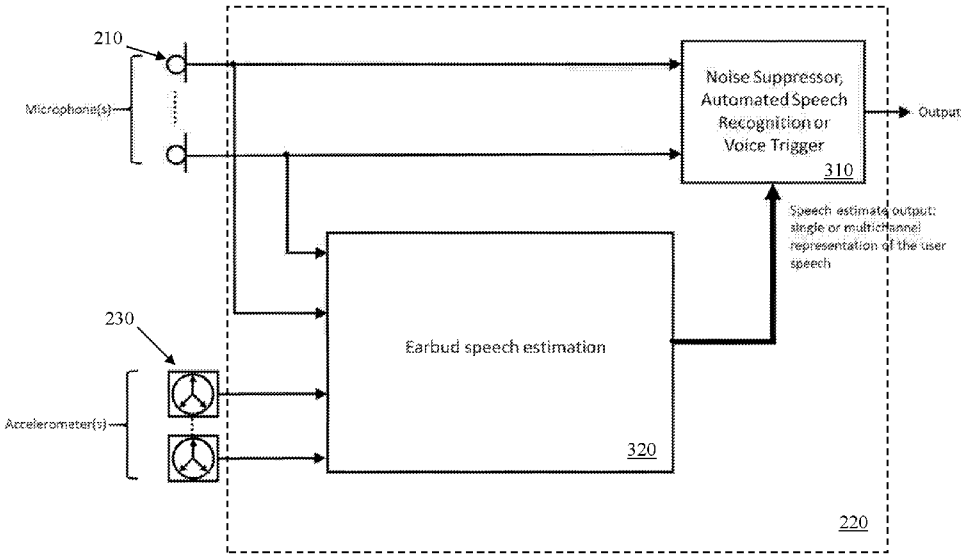

Fig. 3a

(57) Abstract: Embodiments of the invention determine a speech estimate using a bone conduction sensor or accelerometer, without employing voice activity detection gating of speech estimation. Speech estimation is based either exclusively on the bone conduction signal, or is performed in combination with a microphone signal. The speech estimate is then used to condition an output signal of the microphone. There are multiple use cases for speech processing in audio devices.

WO 2018/229503 A1

[Continued on next page]

(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).

Published:

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

EARBUD SPEECH ESTIMATION

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of United States Provisional Patent Application No. 62/520,713 filed 16 June 2017, which is incorporated herein by reference.

FIELD OF THE INVENTION

[0002] The present invention relates to an earbud headset configured to perform speech estimation, for functions such as speech capture, and in particular the present invention relates to earbud speech estimation based upon a bone conduction sensor signal.

BACKGROUND OF THE INVENTION

[0003] Headsets are a popular way for a user to listen to music or audio privately, or to make a hands-free phone call, or to deliver voice commands to a voice recognition system. A wide range of headset form factors, i.e. types of headsets, are available, including earbuds. The in-ear position of an earbud when in use presents particular challenges to this form factor. The in-ear position of an earbud heavily constrains the geometry of the device and significantly limits the ability to position microphones widely apart, as is required for functions such as beam forming or sidelobe cancellation. Additionally, for wireless earbuds the small form factor places significant limitations on battery size and thus the power budget. Moreover, the anatomy of the ear canal and pinna somewhat occludes the acoustic signal path from the user's mouth to microphones of the earbud when placed within the ear canal, increasing the difficulty of the task of differentiating the user's own voice from the voices of other people nearby.

[0004] Speech capture generally refers to the situation where the headset user's voice is captured and any surrounding noise, including the voices of other people, is minimised. Common scenarios for this use case are when the user is making a voice call, or interacting with a speech recognition system. Both of these scenarios place stringent requirements on the underlying algorithms. For voice calls, telephony standards and user

requirements demand that high levels of noise reduction are achieved with excellent sound quality. Similarly, speech recognition systems typically require the audio signal to have minimal modification, while removing as much noise as possible. Numerous signal processing algorithms exist in which it is important for operation of the algorithm to change, depending on whether or not the user is speaking. Voice activity detection, being the processing of an input signal to determine the presence or absence of speech in the signal, is thus an important aspect of voice capture and other such signal processing algorithms. However, even in larger headsets such as booms, pendants, and supra-aural headsets, it is very difficult to reliably ignore speech from other persons who are positioned within a beam of a beamformer of the device, with the consequence that such other persons' speech can corrupt the process of voice capture of the user only. These and other aspects of voice capture are particularly difficult to effect with earbuds, including for the reason that earbuds do not have a microphone positioned near the user's mouth and thus do not benefit from the significantly improved signal to noise ratio resulting from such microphone positioning.

[0005] Any discussion of documents, acts, materials, devices, articles or the like which has been included in the present specification is solely for the purpose of providing a context for the present invention. It is not to be taken as an admission that any or all of these matters form part of the prior art base or were common general knowledge in the field relevant to the present invention as it existed before the priority date of each claim of this application.

[0006] Throughout this specification the word "comprise", or variations such as "comprises" or "comprising", will be understood to imply the inclusion of a stated element, integer or step, or group of elements, integers or steps, but not the exclusion of any other element, integer or step, or group of elements, integers or steps.

[0007] In this specification, a statement that an element may be "at least one of" a list of options is to be understood that the element may be any one of the listed options, or may be any combination of two or more of the listed options.

SUMMARY OF THE INVENTION

[0008] According to a first aspect the present invention provides a signal processing device for earbud speech estimation, the device comprising:

at least one input for receiving a microphone signal from a microphone of an earbud;

at least one input for receiving a bone conduction sensor signal from a bone conduction sensor of an earbud;

a processor configured to determine from the bone conduction sensor signal at least one characteristic of speech of a user of the earbud, the at least one characteristic being a non-binary variable, the processor further configured to derive from the at least one characteristic of speech at least one signal conditioning parameter; and the processor further configured to use the at least one signal conditioning parameter to condition the microphone signal.

[0009] According to a second aspect the present invention provides a method of conditioning an earbud microphone signal, the method comprising:

receiving a bone conduction sensor signal from a bone conduction sensor of an earbud;

receiving a microphone signal from a microphone of the earbud;

determining from the bone conduction sensor signal at least one characteristic of speech of a user of the earbud, the at least one characteristic being a non-binary variable;

deriving from the at least one characteristic of speech at least one signal conditioning parameter; and

using the at least one signal conditioning parameter to condition the output signal from the microphone.

[0010] According to a third aspect the present invention provides a non-transitory computer readable medium for conditioning an earbud microphone signal, comprising instructions which, when executed by one or more processors, causes performance of the following:

receiving a bone conduction sensor signal from a bone conduction sensor of an earbud;

receiving a microphone signal from a microphone of the earbud;

determining from the bone conduction sensor signal at least one characteristic of speech of a user of the earbud, the at least one characteristic being a non-binary variable;

deriving from the at least one characteristic of speech at least one signal conditioning parameter; and

using the at least one signal conditioning parameter to condition the output signal from the microphone.

[0011] In some embodiments the earbud is a wireless earbud.

[0012] The non-binary variable characteristic of speech determined by the processor from the bone conduction sensor signal in some embodiments is a speech estimate derived from the bone conduction sensor signal. The processor may in some embodiments be configured such that the conditioning of the microphone signal comprises non-stationary noise reduction controlled by the speech estimate derived from the bone conduction sensor signal. The non-stationary noise reduction may in some embodiments be further controlled by a speech estimate derived from the microphone signal.

[0013] The processor may in some embodiments be configured such that the non-binary variable characteristic of speech determined from the bone conduction sensor signal is a speech level of the bone conduction sensor signal.

[0014] The processor may in some embodiments be configured such that the non-binary variable characteristic of speech determined from the bone conduction sensor signal is an observed spectrum of the bone conduction sensor signal.

[0015] The processor may in some embodiments be configured such that the non-binary variable characteristic of speech determined from the bone conduction sensor signal is a parametric representation of the spectral envelope of the bone conduction sensor signal.

[0016] The processor may in some embodiments be configured such that the parametric representation of the spectral envelope of the bone conduction sensor signal comprises at least one of: linear prediction cepstral coefficients, autoregressive coefficients, and line spectral frequencies, for example to model the human vocal tract in order to derive the speech envelope.

[0017] The processor may in some embodiments be configured such that the non-binary variable characteristic of speech determined from the bone conduction sensor signal is a non-parametric representation of the spectral envelope of the bone conduction sensor signal, such as mel-frequency cepstral coefficients (MFCCs) derived from models of human sound perception, or log-spaced spectral magnitudes derived from a short time Fourier transform which is a preferred method.

[0018] The processor may in some embodiments be configured such that the conditioning of the output signal from the microphone occurs irrespective of voice activity.

[0019] The processor may in some embodiments be configured such that the at least one signal conditioning parameter comprises band-specific gains derived from the bone conduction sensor signal, and wherein the conditioning of the microphone signal comprises applying the band-specific gains to the microphone signal.

[0020] The processor may in some embodiments be configured such that the conditioning of the microphone signal comprises applying a Kalman filter process in which the bone conduction sensor signal acts a priori to a speech estimation process. A speech estimate may in some embodiments be derived from the bone conduction sensor signal and be used to modify a decision-directed weighting factor for *a priori* SNR estimation. A speech estimate derived from the bone conduction sensor signal may in some embodiments be used to inform an update step in a casual recursive speech enhancement (CRSE).

[0021] The non-binary variable characteristic of speech determined by the processor from the bone conduction sensor signal may in some embodiments be a signal to noise ratio of the bone conduction sensor signal.

[0022] The processor may in some embodiments be configured such that, other than the bone conduction sensor signal being a basis for determination of the at least one characteristic of speech, no component of the bone conduction sensor signal is passed to a signal output of the earbud.

[0023] The processor may in some embodiments be configured such that, before the non-binary variable characteristic of speech is determined from the bone conduction sensor signal, the bone conduction sensor signal is corrected for observed conditions. The processor may in some embodiments be configured such that the bone conduction sensor signal is corrected for phoneme. The processor may in some embodiments be configured such that the bone conduction sensor signal is corrected for bone conduction coupling. The processor may in some embodiments be configured such that the bone conduction sensor signal is corrected for bandwidth. The processor may in some embodiments be configured such that the bone conduction sensor signal is corrected for distortion. The processor may in some embodiments be configured to perform the correction of the bone conduction sensor signal by applying a mapping process. The mapping process may in some embodiments comprise a linear mapping involving a series of corrections associated with each spectral bin of the bone conduction sensor signal. For example, the corrections may comprise a multiplier and offset applied to the respective spectral bin value of the bone conduction sensor signal. The processor may in some embodiments be configured to perform the correction of the bone conduction sensor signal by applying offline learning.

[0024] The processor may in some embodiments be configured such that the conditioning of the microphone signal is based only upon the non-binary variable characteristic of speech determined from the bone conduction sensor signal.

[0025] The bone conduction sensor may in some embodiments comprise an accelerometer, which in use is coupled to a surface of the user's ear canal or concha, to detect bone conducted signals from the user's speech.

[0026] The bone conduction sensor may in some embodiments be comprise an in-ear microphone which in use is positioned to detect acoustic sounds arising within the ear canal as a result of bone conduction of the user's speech. The accelerometer and the in-ear microphone may in some embodiments both be used to detect at least one characteristic of speech of the user.

[0027] The processor may in some embodiments be configured to apply at least one matched filter to the bone conduction sensor signal, the matched filter being configured to match the user's speech in the bone conduction sensor signal to the user's speech in the microphone signal. The matched filter may in some embodiments have a design which is based on a training set.

[0028] The processor may in some embodiments be configured to condition the microphone signal unilaterally, without input from any contralateral sensor on an opposite ear of the user.

[0029] An earbud is defined herein as an audio headset device, whether wired or wireless, which in use is supported only or substantially by the ear upon which it is placed, and which comprises an earbud body which in use resides substantially or wholly within the ear canal and/or concha of the pinna.

BRIEF DESCRIPTION OF THE DRAWINGS

[0030] An example of the invention will now be described with reference to the accompanying drawings, in which:

Fig. 1 illustrates the use of wireless earbuds for telephony and/or audio playback;

Fig. 2 is a system schematic of an earbud in accordance with one embodiment of the invention;

Figs. 3a and 3b are detailed system schematics of the earbud of Figure 2;

Fig. 4 is a flow diagram for the earbud speech estimation process of the embodiment of Fig. 3;

Fig. 5 illustrates a noise suppressor for telephony in accordance with another embodiment of the invention;

Fig. 6 illustrates an embodiment comprising a speech estimator that uses a statistical model based estimation process;

Fig. 7 illustrates a mic-accelerometer mixing approach which is based on mixing factors using SNR estimates;

Fig. 8 illustrates the configuration of another embodiment of the invention;

Fig. 9 illustrates an embodiment applying speech estimation from a bone conduction sensor signal to the telephony use case; and

Fig. 10 shows objective Mean Opinion Score (MOS) results for one embodiment of the invention.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

[0031] **Fig. 1** illustrates the use of wireless earbuds for telephony and/or audio playback. Device 110, which may be a smartphone or audio player or the like, communicates with bilateral wireless earbuds 120, 130. For illustrative purposes earbuds 120, 130 are shown outside the ear however in use each earbud is placed so that the body of the earbud resides substantially or wholly within the concha and/or ear canal of the respective ear. Earbuds 120, 130 may each take any suitable form to comfortably fit upon or within, and be supported by, the ear of the user. In some embodiments within the scope of the present invention the body of the earbud may be further supported by a hook or support member extending beyond the concha such as partly or completely around the outside of the respective pinna.

[0032] **Fig. 2** illustrates the system of earbud 120. Earbud 130 may be similarly configured and is not described separately. A microphone 210 is positioned on earbud 120 so as to receive external acoustic signals when the earbud is in place. A plurality of microphones may be provided, for example in order to enable beamforming noise reduction to be undertaken by the earbud 120, however the small size of earbud 120 places a difficult limitation on the maximum microphone spacing which can be

implemented, and the positioning of the earbud in a position where sound is partly occluded or diffused by the pinna are factors which both limit the efficacy of beamforming, as compared to say a boom-mounted microphone.

[0033] The microphone signal from microphone 210 is passed to a suitable processor 220 of earbud 120. Due to the size of earbud 120 limited battery power is available which dictates that processor 220 executes only low power and computationally simple audio processing functions.

[0034] Earbud 120 further comprises an accelerometer 230 which is mounted upon earbud 120 in a location which is inserted into the ear canal and pressed against a wall of the ear canal in use, or as appropriate accelerometer 230 may be mounted within a body of the earbud 120 so as to be mechanically coupled to a wall of the ear canal. Accelerometer 230 is thereby configured to detect bone conducted signals, and in particular the user's own speech as conducted by the bone and tissue interposed between the vocal tract and the ear canal. Such signals are referred to herein as bone conducted signals, even though acoustic conduction may occur through other body tissue and may partly contribute to the signal sensed by the bone conduction sensor 230.

[0035] The bone conduction sensor could in alternative embodiments be coupled to the concha or mounted upon any part of the headset body that reliably contacts the ear within the ear canal or concha. The use of an earbud allows for reliable direct contact with the ear canal and therefore a mechanical coupling to the vibration model of bone conducted speech as measured at the wall of the ear canal. This is in contrast to the external temple, cheek or skull, where a mobile device such as a phone might make contact. The present invention recognises that a bone conducted speech model derived from parts of the anatomy outside the ear produces a signal that is significantly less reliable for speech estimation as compared to described embodiments of this invention. The present invention recognises that use of a bone conduction sensor in a wireless earbud is sufficient to perform speech estimation. This is because, unlike a handset or a headset outside the ear, the nature of the bone conduction sensor signal from wireless earbuds is largely static with regard to the user fit, user actions and user movements. For

example the present invention recognises that no compensation of the bone conduction sensor is required for fit or proximity. Thus, selection of the ear canal or concha as the location for the bone conduction sensor is a key enabler for the present invention. In turn, the present invention then turns to deriving a transformation of that signal that best identifies the temporal and spectral characteristics of user speech.

[0036] The device 120 is a wireless earbud. This is important as the accessory cable attached to wired personal audio devices is a significant source of external vibration to the bone conduction sensor 230. The accessory cable also increases the effective mass of the device 120 which can damp vibrations of the ear canal due to bone conducted speech. Eliminating the cable also reduces the need for a compliant medium in which to house the bone conduction sensor 230. The reduced weight increases compliance with the ear canal vibration due to bone conducted speech. Therefore in wireless embodiments of the invention there is no or vastly reduced restrictions on placement of the bone conduction sensor 230. The only requirement is that sensor 230 makes rigid contact with the external housing of the earbud 120. Embodiments thus may include mounting the sensor 230 on a printed circuit board (PCB) inside the earbud housing or to a BTE module coupled to the earbud kernel via a rigid rod.

[0037] The position of the primary voice microphone 210 is generally close to the ear in wireless earbuds. It is therefore relatively distant from the user's mouth and consequently suffers from a low signal to noise ratio (SNR). This is in contrast to a handset or pendant type headset, in which the primary voice microphone is much closer to the mouth, and in which differences in how the user holds the phone/pendant can give rise to a wide range of SNR. In the present embodiment the SNR on the primary voice microphone 210 for a given environmental noise level is not so variable as the geometry between the user's mouth and the ear containing the earbud is fixed. Therefore the ratio between the speech level on the primary voice microphone 210 and the speech level on the bone conduction sensor 230 are known *a priori* and the present invention therefore recognises that this is in part useful for determining the relationship between the true speech estimate and the bone conduction sensor signal.

[0038] The sufficient condition of contact between the bone conduction sensor 230 and the ear canal is due to the weight of the ear bud 120 being small enough that the force of the vibration due to speech exceeds the minimum sensitivity of commercial accelerometers 230. This is in contrast to an external headset or phone handset which has a large mass which prevents bone conducted vibrations from easily coupling to the device.

[0039] Processor 220 is a signal processing device configured to determine from the bone conduction sensor signal from accelerometer 230 at least one characteristic of speech of a user of the earbud 120, derive from the at least one characteristic of speech at least one signal conditioning parameter; and the processor 220 is further configured to use the at least one signal conditioning parameter to condition the microphone signal from microphone 210 and wirelessly deliver the conditioned signal to master device 110 for use as the transmitted signal of a voice call and/or for use in automatic speech recognition (ASR). Communications between earbud 120 and master device 110 may for example be undertaken by way of low energy Bluetooth. Alternative embodiments may utilise wired earbuds and communicate by wire, albeit with the disadvantages discussed elsewhere herein. Speaker 240 is configured to play back acoustic signals into the ear canal of the user, such as a receive signal of a voice call.

[0040] Notably, the present embodiment provides for noise reduction to be applied in a controlled gradated manner, and not in a binary on-off manner, based upon a speech estimation derived from the bone conduction sensor signal, on a headset form factor comprising a wireless earbud provided with at least one microphone and at least one accelerometer. In particular, in contrast to the binary process of voice activity detection, speech estimation involves the estimation of spectral amplitudes or signal peak frequencies and the application of suitable processing to improve speech quality. Indeed some embodiments of the present invention may apply speech estimation based on the bone conduction sensor signal in the absence of any voice activity detection and microphone signal gating step whatsoever.

[0041] Accurate speech estimates can lead to better performance on a range of speech enhancement metrics. Voice activity detection (VAD) is one way of improving the speech estimate but inherently relies on the imperfect notion of identifying in a binary manner the presence or absence of speech in noisy signals. The present embodiment recognises that the accelerometer 230 can capture a suitable noise-free speech estimate that can be derived and used to drive speech enhancement directly, without relying on a binary indicator of speech or noise presence. A number of solutions follow from this recognition.

[0042] **Figs. 3a and 3b** illustrate in greater detail the configuration of processor 220 within the system of earbud 120, in accordance with one embodiment of the invention. The embodiment of Figures 3a and 3b recognises that in moderate signal to noise ratio (SNR) conditions, improved non-stationary noise reduction can be achieved with speech estimates alone, without VAD. This is distinct from approaches in which voice activity detection is used to discriminate between the presence of speech and the absence of speech, and a discrete binary decision signal from the VAD is used to gate, i.e. turn on and off, a noise suppressor acting on an audio signal. The embodiment of Figure 3 recognises that the accelerometer signal or some signal derived from it may be relied upon to obtain sufficiently accurate speech estimates, even in acoustic conditions where accurate speech estimations cannot be obtained from the microphone signal. Omission of the VAD in such embodiments contributes to minimising the computational burden on the earbud processor 220.

[0043] In more detail, in Figure 3 the microphone signal from microphone 210 is conditioned by a noise suppressor 310, and then passed to an output, such as for wireless communication to device 110. The noise suppressor 310 is continually controlled by speech estimation / characterisation module 320, without any on-off gating by any VAD. Speech estimation / characterisation module 320 takes inputs from accelerometer 230, and optionally also from other accelerometers, microphone 210, and/or other microphones.

[0044] The selection of an accelerometer 230 as the bone conduction sensor in such embodiments is particularly useful because the noise floor in commercial accelerometers is, as a first approximation, spectrally flat. These devices are acoustically transparent up to the resonant frequency and so display no signal due to environmental noise. The noise distribution of the sensor 230 can therefore be updated *a priori* to the speech estimation process. This is an important difference as it permits modelling of the temporal and spectral nature of the true speech signal without interference by the dynamics of a complex noise model. Experiments show that even tethered (wired) earbuds have a complex noise model due to short term changes in the temporal and spectral dynamics of noise due to events such as cable bounce. Corrections to the bone conduction spectral envelope in wireless earbud 120 are not required as a matched signal is not a requirement for the design of a conditioning parameter.

[0045] Speech estimation 320 is performed on the basis of certain signal guarantees in the microphone(s) 210 and accelerometers 230, as are guaranteed in the wireless earbud use case in particular. However, corrections to the bone conduction spectral envelope in an earbud may be performed to weight feature importance but a matched signal is not a requirement for the design of a conditioning parameter. Sensor non-idealities and non-linearities in the bone conduction model of the ear canal are other reasons a correction may be applied.

[0046] In particular, embodiments employing multiple bone conduction sensors 230 in the ear are proposed to be configured so as to exploit orthogonal modes of vibration arising from bone conducted speech in the ear canal in order to extract more information about the user speech. Importantly, the bone conducted signal couples reliably into the sensors within the scope of wireless earbuds, unlike wired earbuds to an extent, and unlike headsets outside the ear. In such embodiments the problem of capturing various modalities of bone conducted speech in the ear canal is solved by the use of multiple bone conduction devices arranged orthogonally in the earbud housing, or by a single bone conduction device with independent orthogonal axes.

[0047] The signal from accelerometer 230 is high pass filtered and then used by module 320 to determine a speech estimate output which may comprise a single or multichannel representation of the user speech, such as a clean speech estimate, the *a priori* SNR, and/or model coefficients.

[0048] Notably, the configuration of Figure 3 omits any voice activity detection (VAD). Numerous methods of speech enhancement rely on various estimates of the speech signal, and become challenging when microphone speech signals become degraded by environmental noise. The accuracy of these estimates generally diminishes with the level of environmental noise. The uses for speech estimates include wind noise suppression, *a priori* SNR estimation for noise suppression, biasing of the gain function for noise suppression, beamforming adaption (blocking matrix update), adaption control for acoustic echo cancellation, *a priori* speech to echo estimation for echo suppression, adaptive thresholding for VAD (level difference and cross-correlation), and adaptive windowing for stationary noise estimates (minima controlled recursive averaging; MCRA).

[0049] The processing of the bone conduction sensor 230 and consequent conditioning occurs irrespective of speech activity in an accelerometer signal in this embodiment of the invention. It is therefore not dependent on either a speech detection process or noise modelling (VAD) process in deriving the speech estimate for a noise reduction process. The noise statistics of an accelerometer sensor 230 measuring ear canal vibrations in a wireless earbud 120 have a well-defined distribution unlike the handset use case. The present invention recognises that this justifies a continuous speech estimation based on the signal from accelerometer 230. Although the microphone 210 SNR will be lower in an earbud due to distance of the microphone 210 from the mouth, the distribution of speech samples will have a lower variance than that of a handset or pendant due to the fixed position of the earbud and microphone 210 relative to the mouth. This collectively forms the *a priori* knowledge of the user speech signal to be used in the conditioning parameter design and speech estimation processes 320.

[0050] The embodiment of Figure 3 recognises that speech estimation using a microphone and bone conduction sensor can improve speech estimation for such purposes. The speech estimate may be derived from the bone conduction sensor (e.g. accelerometer 230) or a combination of both bone conduction sensor(s) 230 and microphone(s) 210. The speech estimate from the bone conduction sensor 230 may comprise any combination of signals from separate axes of a single device. The speech estimate may be derived from time domain or frequency domain signals. By undertaking the processing within the earbud 120 rather than in master device 110, the processor 220 can be configured at a time of manufacture or configuration with certainty that the described processes have access to all of the appropriate signals and are based on precise knowledge of the earbud geometry.

[0051] Before the non-binary variable characteristic of speech is determined from the bone conduction sensor signal, the bone conduction sensor signal is corrected for observed conditions, and for example the bone conduction sensors signal may be corrected for phoneme, sensor bandwidth and/or distortion. The correction may involve a linear mapping which undertakes a series of corrections associated with each spectral bin, such as applying a multiplier and offset to each bin value.

[0052] The speech estimates may be derived at 320 from the bone conduction sensor 230 by any of the following techniques: exponential filtering of signals (leaky integrator); gain function of signal values; fixed matching filter (FIR or spectral gain function); adaptive matching (LMS or input signal driven adaptation); mapping function (codebook); and using second order statistics to update an estimation routine. In addition, speech estimates may be derived from different signals for different amplitudes of the input signals, or other metric of the input signals such as noise levels. For example, the accelerometer 230 noise floor is much higher than the microphone 210 noise floor, and so below some nominal level the accelerometer information may no longer be as useful and the speech estimate can transition to a microphone-derived signal. The speech estimates as a function of input signals may be piecewise or continuous over transition regions. Estimation may vary in method and may rely on different signals with each region of the transfer curve. This will be determined by the use case, such as a noise

suppression long term SNR estimate, noise suppression *a priori* SNR reduction, and gain back-off.

[0053] **Fig. 3b** provides more detail of the earbud speech estimation process 320 of Fig. 3a. **Fig. 4** is a flow diagram for the earbud speech estimation process.

[0054] Notably, Figs. 3a and 3b describe a speech estimator 320 conditioned on the bone conduction speech signal from 230. This estimation may take the form of a time and/or frequency domain signal representative of the user speech signal. This is distinct from a clean speech signal that may be the result of an application of this estimator 320.

[0055] A noise suppressor for telephony as shown in **Fig. 5** may use the estimator in producing a clean speech signal that will be transferred across a telephony network to a remote recipient. Examples of noise suppressors include Spectral Subtraction, Wiener Filtering and Statistical Model Methods.

[0056] An example of an embodiment of the speech estimator that uses a statistical model based estimation process is shown in **Fig. 6**. The air conducted microphone speech estimate, the bone conducted speech estimate and SNR are separately derived from a causal recursive speech enhancement process. *A priori* SNR estimates from each process are then combined to derive mixing coefficients that condition the user speech estimates to arrive at a final speech estimator. It is important to note that neither the microphone nor the accelerometer sensor signals are used to derive a noise model in this process. Instead the information content within the signals as influenced by the wireless earbud form factor allow a direct speech estimation process.

[0057] In another example the application may be in producing a signal representative of a latent representation of speech suitable for an Automated Speech Recognition (ASR) system. In this case the latent representation of the clean speech is derived from a transformation of the speech estimator.

[0058] The distinction of this approach is recognised in the exploitation of the temporal and spectral dynamics of the bone conduction signal in the presence of a stationary noise signal to derive a speech model. This is in contrast to the exploitation of the same dynamics for speech detection which find widespread application in the field of voice activity detectors.

[0059] Corrections to the bone conduction spectral envelope in an earbud may be performed to weight feature importance but a matched signal is not a requirement for the design of a conditioning parameter.

[0060] The approach to derive a speech estimator, in contrast to a speech detector (VAD), using the bone conduction sensor can be further elaborated upon within the context of this invention. Traditionally the quality of noise suppressors is dependent on estimates of the noise spectrum. The noise spectrum is typically derived from measurement during speech gaps with a binary decision device such as a VAD. VADs tend to perform poorly in low SNR conditions resulting in errors in the gain function that give rise to the familiar undesirable 'musical noise' phenomena. Alternatively, noise estimates may be obtained by assuming certain statistical properties of the noise signal however, noise statistics of realistic environments can deviate from these assumptions. Since the accuracy of the gain function is highly dependent on the SNR estimate this means that, in the absence of accurate noise statistics, SNR estimation can exploit knowledge of the speech estimate.

[0061] The present invention does not use the bone conduction sensor in the process of building a noise model. Therefore construction of a noise model does not require a voice activity detector (VAD) derived from the bone conduction sensor. This is an important contrast with other proposals to use a bone conduction sensor as a substitute for a microphone, as in such alternative proposals typically the noise model must be accurately modelled for performing speech enhancement and therefore the bone conduction sensor is instrumental in deriving that model.

[0062] The bone conduction sensor in the present invention is for deriving one or more conditioning parameters for the microphone speech envelope, and is inherently bone conduction VAD-free. The nature of wireless earbuds as previously discussed avoids the need to consider a complex noise model introduced by the bone conduction sensor. In contrast the underlying assumption of the bone conduction sensor in the earbud is that the bone conduction sensor signal representative of speech contains the temporal and spectral content sufficient for deriving a non-binary signal representative of user speech. Thus, the present invention recognises that in the earbud use case the clean speech estimate is not dependent on a bone conduction derived noise estimate. Indeed, the inclusion of a noise model is optional when forming the clean speech estimate although in some instances it may improve the clean speech estimate.

[0063] In one embodiment (FIG 6) the speech model from the noisy microphone may be refined with a causal recursive speech estimator which requires an estimate of the noise variance. This is typically a minimal-tracking or time-recursive averaging algorithm and such estimation is performed in the absence of any specific speech detection. Further, the power spectrum of the bone conduction sensor is by virtue of its representation of ear canal vibration, treated as a prior of the user speech. It need not undergo a transformation to approximate a clean speech microphone signal. In this case it is treated as S_{bc}, a bone conduction speech estimate, rather than a clean speech estimate conditioned on the bone conduction sensor i.e. $\hat{S}_{x|bc}$. In some embodiments S_{bc} may be further refined, for example by the aforementioned CRSE process. Thus, the present embodiments use the bone conduction sensor signal as a prior for clean speech estimation. Notably, these embodiments do not use an offline process to derive a bone conduction to clean air conduction microphone transformation, nor do these embodiments use such as resultant signal as a conditional estimate. Some embodiments of the invention may apply corrections for some non-idealities but, importantly, it is not necessary to add prior information to the signal from any offline process. The present invention recognises that it is possible to do so because the bone conduction sensor signal as a prior is sufficient because of the earbud use case.

[0064] **Fig. 7** illustrates a mic-accelerometer mixing approach which is based on mixing factors using SNR estimates and provides a means to combine *a priori* SNR estimates from the mic and accelerometer (BC sensor). This may be particularly suitable in low SNR environments where the best speech estimate in terms of the SNR estimate is being used. The clean speech estimate and *a priori* SNR estimates derived from the bone conduction sensor signal are thus an application of the bone conduction sensor signal –controlled speech estimation technique in accordance with the present invention. It is to be noted in Figure 7 that the mixing is achieved without use of a VAD. For example, in one approach of mixing the combiner 730 mixes noisy microphone (mic) and bone conduction sensor (accel) signals according to mixing factors α and β derived from respective *a priori* (apr) SNR estimates as follows:

$$\hat{x}_{\Sigma} = \alpha\hat{x}_{mic} + \beta\hat{x}_{accel}$$

$$\alpha = \frac{\widehat{SNR}_{mic}^{apr}}{\widehat{SNR}_{mic}^{apr} + \widehat{SNR}_{accel}^{apr}}$$

$$\beta = \frac{\widehat{SNR}_{accel}^{apr}}{\widehat{SNR}_{mic}^{apr} + \widehat{SNR}_{accel}^{apr}}$$

and then a second stage noise reduction is performed on this mixed signal.

[0065] This is in contrast to using a VAD to derive noise estimates and to subsequently determine mixing ratios.

[0066] Further embodiments of the present invention may enlarge upon this idea by discarding speech estimates from the speech enhancement blocks 710, 720, instead mixing the noisy signals from SNR estimates and performing a second-stage noise reduction.

[0067] **Fig. 8** illustrates the configuration of processor 220 within the system of earbud 120, in accordance with another embodiment of the invention. Elements of Figure 8 not described are as for Figure 3. However, in the embodiment of Figure 8 the speech estimate output by the speech estimation / characterisation module is delivered not only to the noise suppressor but also to a secondary output path for use by other modules which may for example be within the earbud 120 or the master device 110, and

for example could include an automatic speech recognition (ASR) module or could be a voice-triggered module. Design of an appropriate gain function takes place inside the noise suppression model and relies on the conditioned speech estimate of the microphone signal.

[0068] **Fig. 9** illustrates a further embodiment in accordance with the present invention, illustrating the application of the speech estimation from the bone conduction sensor signal to the telephony use case.

[0069] Embodiments of the present invention note that, despite the poor frequency response of in-ear accelerometers as compared to microphones and even as compared to temple mounted bone sensors or the like, it is nevertheless possible to not only use in-ear accelerometer signals for speech estimation but moreover it is recognised that in-ear accelerometer signals may be used for gradated or non-binary control of speech estimation, such as by controlling non-stationary noise reduction in a multi-stepped or gradated manner. In more detail, the low pass frequency response of earbud inertial sensors, and relatively poor sensitivity, are limitations of the bone conduction model at the outer ear canal. Bone conduction sensors for vibration are typically magnetic type and mounted to other parts of the head such as the temporal bone or mastoid bone, often utilising a spring force of a headband or the like to maintain a firm contact. Such mounting locations and techniques however are somewhat incongruent with headsets for audio applications and not compatible with preferred headset form factors. The present invention, in utilising an inertial sensor of an earbud, is beneficial in conforming to a preferred headset form factor.

[0070] The speech spectral envelope in the present embodiments is not a convex combination of microphone signal, noise model and bone conduction signal. This is not practical given the spectral nature of the accelerometer signal used in one of our embodiments since the bone conduction model of speech in the ear canal limits the observable frequency range. Bone conduction models based on other parts of the body can exploit modes of high frequency radiation in excess of 1 kHz. Estimating a time-frequency model of speech in the ear canal is therefore a different problem as the present

inventors have discovered that the observable frequency range of ear canal bone conduction signals is typically below 1 kHz. The present inventors have shown however that temporal and spectral information available from the accelerometer even in such a limited band nevertheless adds information about the nature of the true clean speech that can inform the noise reduction process in a useful way.

[0071] Fig. 10 shows objective Mean Opinion Score (MOS) results for the embodiment of Fig. 9, showing the improvement when the *a priori* speech envelope from the microphone 210 is conditioned with a parameter(s) derived from the bone conduction sensor 230 spectral envelope. The measurements are performed in a number of different stationary and non-stationary noise types using the 3Quest methodology to obtain speech MOS (S-MOS) and noise MOS (N-MOS) values.

[0072] While in other applications such as handsets bone conduction and microphone spectral estimates in the combined estimates have time and frequency contribution that may fall to zero if the handset use case forces either sensor signal quality to be very poor, this is not the case in the wireless earbud application of the present embodiments. In contrast the *a priori* speech estimates of the microphone 210 and accelerometer 230 in the earbud form factor can be combined in a continuous way. For example, provided the earbud 120 is being worn by the user, the accelerometer sensor model will always provide a signal representative of user speech to the conditioning parameter design process. As such, the microphone speech estimate is continuously being conditioned by this parameter.

[0073] While the described embodiments provide for the speech estimation / characterisation 320 module and the noise suppressor module 310 to reside within earbud 120, alternative embodiments may instead or additionally provide for such functionality to be provided by master device 110. Such embodiments may thus utilise the significantly greater processing capabilities and power budget of master device 110 as compared to earbuds 120, 130.

[0074] Earbud 120 may further comprise other elements not shown such as further digital signal processor(s), flash memory, microcontrollers, Bluetooth radio chip or equivalent, and the like.

[0075] The described embodiments utilise accelerometer 230 as the bone conducted signal sensor. However, alternative embodiments may sense bone conducted signals by additionally or alternatively providing one or more in-ear microphones. Such in-ear microphones will, unlike accelerometer 230, receive acoustic reverberations of bone conducted signals which reverberate within the ear canal, and will also receive leakage of external noise into the ear canal past the earbud. However, the present inventors recognise that the earbud provides a significant occlusion of such external noise, and moreover that active noise cancellation (ANC) when employed will further reduce the level of external noise inside the ear canal without significantly reducing the level of bone conducted signal present inside the ear canal, so that an in-ear microphone may indeed capture very useful bone-conducted signals to assist with speech estimation in accordance with the present invention. Additionally, such in-ear microphones may be matched at a hardware level with the external microphone 210, and may capture a broader spectrum than an accelerometer, and thus the use of one or more in-ear microphones may present significantly different implementation challenges to the use of an accelerometer(s).

[0076] The claimed electronic functionality can be implemented by discrete components mounted on a printed circuit board, or by a combination of integrated circuits, or by an application-specific integrated circuit (ASIC). Wireless communications is to be understood as referring to a communications, monitoring, or control system in which electromagnetic or acoustic waves carry a signal through atmospheric or free space rather than along a wire.

[0077] Corresponding reference characters indicate corresponding components throughout the drawings.

[0078] It will be appreciated by persons skilled in the art that numerous variations and/or modifications may be made to the invention as shown in the specific embodiments without departing from the spirit or scope of the invention as broadly described. The present embodiments are, therefore, to be considered in all respects as illustrative and not restrictive.

CLAIMS:

1. A signal processing device for earbud speech estimation, the device comprising:

at least one input for receiving a microphone signal from a microphone of an earbud;

at least one input for receiving a bone conduction sensor signal from a bone conduction sensor of an earbud;

a processor configured to determine from the bone conduction sensor signal at least one characteristic of speech of a user of the earbud, the at least one characteristic being a non-binary variable, the processor further configured to derive from the at least one characteristic of speech at least one signal conditioning parameter; and the processor further configured to use the at least one signal conditioning parameter to condition the microphone signal.

2. The signal processing device according to claim 1, wherein the earbud is a wireless earbud.

3. The signal processing device according to claim 1 or claim 2 wherein the non-binary variable characteristic of speech determined by the processor from the bone conduction sensor signal is a speech estimate derived from the bone conduction sensor signal.

4. The signal processing device according to claim 3 wherein the processor is configured such that the conditioning of the microphone signal comprises non-stationary noise reduction controlled by the speech estimate derived from the bone conduction sensor signal.

5. The signal processing device according to claim 4 wherein non-stationary noise reduction is further controlled by a speech estimate derived from the microphone signal.

6. The signal processing device according to any one of claims 1 to 5 wherein the processor is configured such that the non-binary variable characteristic of speech determined from the bone conduction sensor signal is a speech level of the bone conduction sensor signal.

7. The signal processing device according to any one of claims 1 to 6 wherein the processor is configured such that the non-binary variable characteristic of speech

determined from the bone conduction sensor signal is an observed spectrum of the bone conduction sensor signal.

8. The signal processing device according to claim 7 wherein the processor is configured such that the non-binary variable characteristic of speech determined from the bone conduction sensor signal is a parametric representation of the spectral envelope of the bone conduction sensor signal.

9. The signal processing device according to claim 8 wherein the processor is configured such that the parametric representation of the spectral envelope of the bone conduction sensor signal comprises at least one of: linear prediction cepstral coefficients, autoregressive coefficients, and line spectral frequencies.

10. The signal processing device according to any one of claims 1 to 9 wherein the processor is configured such that the conditioning of the output signal from the microphone occurs irrespective of voice activity.

11. The signal processing device according to any one of claims 1 to 10 wherein the processor is configured such that the at least one signal conditioning parameter comprises band-specific gains derived from the bone conduction sensor signal, and wherein the conditioning of the microphone signal comprises applying the band-specific gains to the microphone signal.

12. The signal processing device according to any one of claims 1 to 11 wherein the processor is configured such that the conditioning of the microphone signal comprises applying a Kalman filter process in which the bone conduction sensor signal acts *a priori* to a speech estimation process.

13. The signal processing device according to claim 12 wherein a speech estimate derived from the bone conduction sensor signal is used to modify a decision-directed weighting factor for *a priori* SNR estimation.

14. The signal processing device according to claim 12 wherein a speech estimate derived from the bone conduction sensor signal is used to inform an update step in a casual recursive speech enhancement (CRSE).

15. The signal processing device according to any one of claims 1 to 14wherein the non-binary variable characteristic of speech determined by the processor from the bone conduction sensor signal is a signal to noise ratio of the bone conduction sensor signal.

16. The signal processing device according to any one of claims 1 to 15 wherein the processor is configured such that, other than the bone conduction sensor signal being a basis for determination of the at least one characteristic of speech, no component of the bone conduction sensor signal is passed to a signal output of the earbud.

17. The signal processing device according to any one of claims 1 to 16 wherein the processor is configured such that, before the non-binary variable characteristic of speech is determined from the bone conduction sensor signal, the bone conduction sensor signal is corrected for observed conditions.

18. The signal processing device according to claim 17 wherein the processor is configured such that the bone conduction sensor signal is corrected for phoneme.

19. The signal processing device according to claim 17 or claim 18 wherein the processor is configured such that the bone conduction sensor signal is corrected for bone conduction coupling.

20. The signal processing device according to any one of claims 17 to 19 wherein the processor is configured such that the bone conduction sensor signal is corrected for bandwidth.

21. The signal processing device according to any one of claims 17 to 20 wherein the processor is configured such that the bone conduction sensor signal is corrected for distortion.

22. The signal processing device according to any one of claims 17 to 21 wherein the processor is configured to perform the correction of the bone conduction sensor signal by applying a mapping process.

23. The signal processing device according to claim 22 wherein the mapping process comprises a linear mapping involving a series of corrections associated with each spectral bin of the bone conduction sensor signal.

24. The signal processing device according to claim 23 wherein the corrections comprise a multiplier and offset applied to the respective spectral bin value of the bone conduction sensor signal

25. The signal processing device according to any one of claims 17 to 24 wherein the processor is configured to perform the correction of the bone conduction sensor signal by applying offline learning.

26. The signal processing device according to any one of claims 1 to 25 wherein the processor is configured such that the conditioning of the microphone signal is based only upon the non-binary variable characteristic of speech determined from the bone conduction sensor signal.

27. The signal processing device according to any one of claims 1 to 26 wherein the bone conduction sensor comprises an accelerometer, which in use is coupled to a surface of the user's ear canal or concha, to detect bone conducted signals from the user's speech.

28. The signal processing device according to any one of claims 1 to 27 wherein the bone conduction sensor comprises an in-ear microphone which in use is positioned to detect acoustic sounds arising within the ear canal as a result of bone conduction of the user's speech.

29. The signal processing device according to claim 27 and claim 28, wherein both the accelerometer and the in-ear microphone are used to detect at least one characteristic of speech of the user.

30. The signal processing device according to any one of claims 1 to 29 wherein the processor is configured to apply at least one matched filter to the bone conduction sensor signal, the matched filter being configured to match the user's speech in the bone conduction sensor signal to the user's speech in the microphone signal.

31. The signal processing device according to claim 30, wherein the at least one matched filter has a design which is based on a training set.

32. The signal processing device according to any one of claims 1 to 31 wherein the processor is configured to condition the microphone signal unilaterally, without input from any contralateral sensor on an opposite ear of the user.

33. A method of conditioning an earbud microphone signal, the method comprising:

receiving a bone conduction sensor signal from a bone conduction sensor of an earbud;

receiving a microphone signal from a microphone of the earbud;

determining from the bone conduction sensor signal at least one characteristic of speech of a user of the earbud, the at least one characteristic being a non-binary variable;

deriving from the at least one characteristic of speech at least one signal conditioning parameter; and

using the at least one signal conditioning parameter to condition the output signal from the microphone.

34. The method of claim 33 wherein the earbud is a wireless earbud.

35. The method according to claim 33 or claim 34 wherein the non-binary variable characteristic of speech determined by the processor from the bone conduction sensor signal is a speech estimate derived from the bone conduction sensor signal.

36. The method according to claim 35 wherein the processor is configured such that the conditioning of the microphone signal comprises non-stationary noise reduction controlled by the speech estimate derived from the bone conduction sensor signal.

37. The method according to claim 36 wherein non-stationary noise reduction is further controlled by a speech estimate derived from the microphone signal.

38. The method according to any one of claims 33 to 37 wherein the processor is configured such that the non-binary variable characteristic of speech determined from the bone conduction sensor signal is a speech level of the bone conduction sensor signal.

39. The method according to any one of claims 33 to 38 wherein the processor is configured such that the non-binary variable characteristic of speech determined from the bone conduction sensor signal is an observed spectrum of the bone conduction sensor signal.

40. The method according to claim 39 wherein the processor is configured such that the non-binary variable characteristic of speech determined from the bone conduction sensor signal is a parametric representation of the spectral envelope of the bone conduction sensor signal.

41. The method according to claim 40 wherein the processor is configured such that the parametric representation of the spectral envelope of the bone conduction sensor signal comprises at least one of: linear prediction cepstral coefficients, autoregressive coefficients, and line spectral frequencies.

42. The method according to any one of claims 33 to 41 wherein the processor is configured such that the conditioning of the output signal from the microphone occurs irrespective of voice activity.

43. The method according to any one of claims 33 to 42 wherein the processor is configured such that the at least one signal conditioning parameter comprises band-specific gains derived from the bone conduction sensor signal, and wherein the conditioning of the microphone signal comprises applying the band-specific gains to the microphone signal.

44. The method according to any one of claims 33 to 43 wherein the processor is configured such that the conditioning of the microphone signal comprises applying a Kalman filter process in which the bone conduction sensor signal acts *a priori* to a speech estimation process.

45. The method according to claim 44 wherein a speech estimate derived from the bone conduction sensor signal is used to modify a decision-directed weighting factor for *a priori* SNR estimation.

46. The method according to claim 44 wherein a speech estimate derived from the bone conduction sensor signal is used to inform an update step in a casual recursive speech enhancement (CRSE).

47. The method according to any one of claims 33 to 46 wherein the non-binary variable characteristic of speech determined by the processor from the bone conduction sensor signal is a signal to noise ratio of the bone conduction sensor signal.

48. The method according to any one of claims 33 to 47 wherein the processor is configured such that, other than the bone conduction sensor signal being a basis for determination of the at least one characteristic of speech, no component of the bone conduction sensor signal is passed to a signal output of the earbud.

49. The method according to any one of claims 33 to 48 wherein the processor is configured such that, before the non-binary variable characteristic of speech is determined from the bone conduction sensor signal, the bone conduction sensor signal is corrected for observed conditions.

50. The method according to claim 49 wherein the processor is configured such that the bone conduction sensor signal is corrected for phoneme.

51. The method according to claim 49 or claim 50 wherein the processor is configured such that the bone conduction sensor signal is corrected for bone conduction coupling.

52. The method according to any one of claims 49 to 51 wherein the processor is configured such that the bone conduction sensor signal is corrected for bandwidth.

53. The method according to any one of claims 49 to 52 wherein the processor is configured such that the bone conduction sensor signal is corrected for distortion.

54. The method according to any one of claims 49 to 53 wherein the processor is configured to perform the correction of the bone conduction sensor signal by applying a mapping process.

55. The method according to claim 54 wherein the mapping process comprises a linear mapping involving a series of corrections associated with each spectral bin of the bone conduction sensor signal.

56. The method according to claim 55 wherein the corrections comprise a multiplier and offset applied to the respective spectral bin value of the bone conduction sensor signal

57. The method according to any one of claims 49 to 56 wherein the processor is configured to perform the correction of the bone conduction sensor signal by applying offline learning.

58. The method according to any one of claims 33 to 57 wherein the processor is configured such that the conditioning of the microphone signal is based only upon the non-binary variable characteristic of speech determined from the bone conduction sensor signal.

59. The method according to any one of claims 33 to 58 wherein the bone conduction sensor comprises an accelerometer, which in use is coupled to a surface of the user's ear canal or concha, to detect bone conducted signals from the user's speech.

60. The method according to any one of claims 33 to 59 wherein the bone conduction sensor comprises an in-ear microphone which in use is positioned to detect acoustic sounds arising within the ear canal as a result of bone conduction of the user's speech.

61. The method according to claim 59 and claim 60, wherein both the accelerometer and the in-ear microphone are used to detect at least one characteristic of speech of the user.

62. The method according to any one of claims 33 to 61 wherein the processor is configured to apply at least one matched filter to the bone conduction sensor signal, the

matched filter being configured to match the user's speech in the bone conduction sensor signal to the user's speech in the microphone signal.

63. The method according to claim 62, wherein the at least one matched filter has a design which is based on a training set.

64. The method according to any one of claims 33 to 63 wherein the processor is configured to condition the microphone signal unilaterally, without input from any contralateral sensor on an opposite ear of the user.

65. A non-transitory computer readable medium for conditioning an earbud microphone signal, comprising instructions which, when executed by one or more processors, causes performance of the following:

receiving a bone conduction sensor signal from a bone conduction sensor of an earbud;

receiving a microphone signal from a microphone of the earbud;

determining from the bone conduction sensor signal at least one characteristic of speech of a user of the earbud, the at least one characteristic being a non-binary variable;

deriving from the at least one characteristic of speech at least one signal conditioning parameter; and

using the at least one signal conditioning parameter to condition the output signal from the microphone.

66. The non-transitory computer readable medium of claim 65 further configured to perform the method of any one of claims 34 to 64.

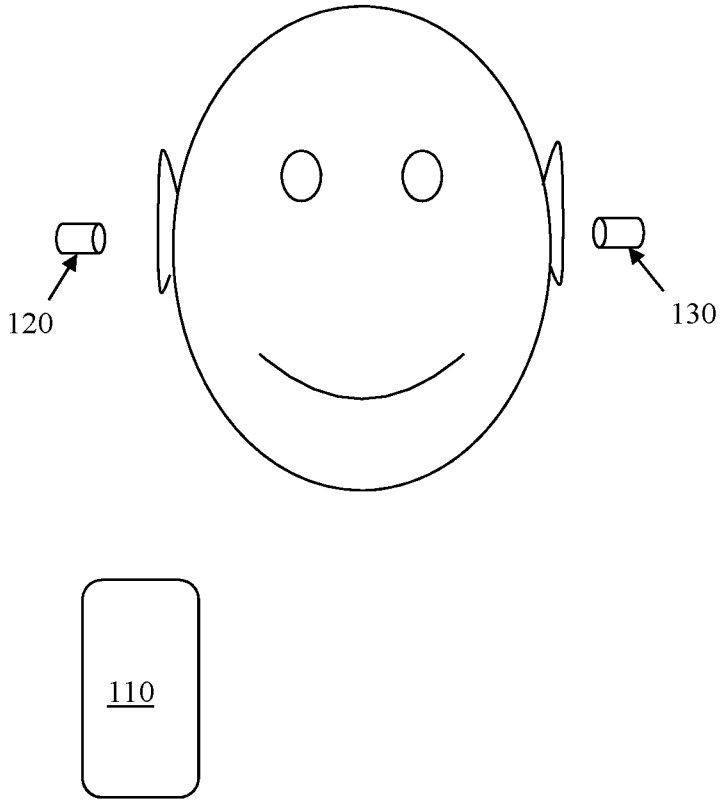

Fig. 1

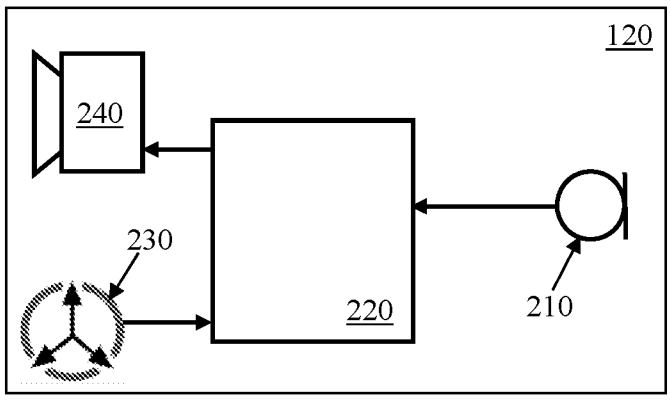

Fig. 2

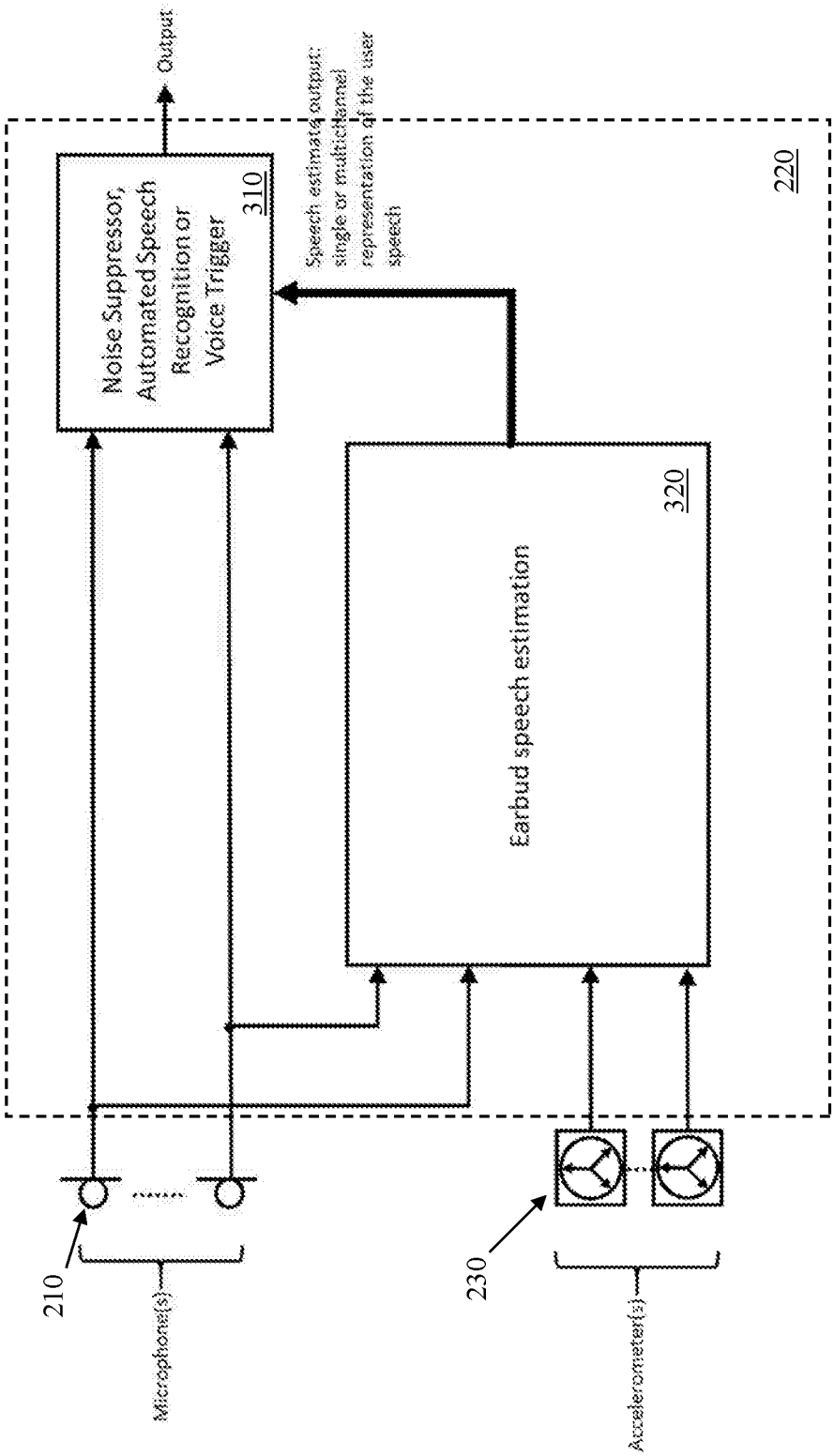
Output
Noise Suppressor, Automated Speech Recognition or Voice Trigger
310
Speech estimate output: single or multichannel representation of the user speech
220
Earbud speech estimation
320
210
Microphone(s)
230
Accelerometer(s)

Fig. 3a

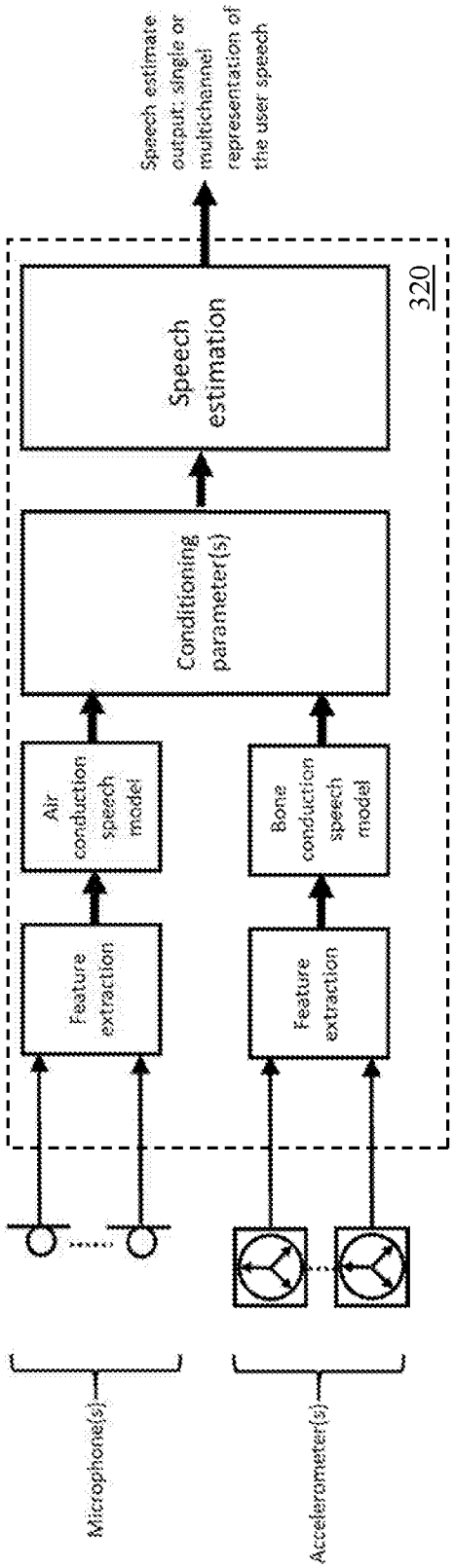
Speech estimate output: single or multichannel representation of the user speech
320
Speech estimation
Conditioning parameter(s)
Air conduction speech model
Bone conduction speech model
Feature extraction
Feature extraction
Microphone(s)
Accelerometer(s)

Fig. 3b

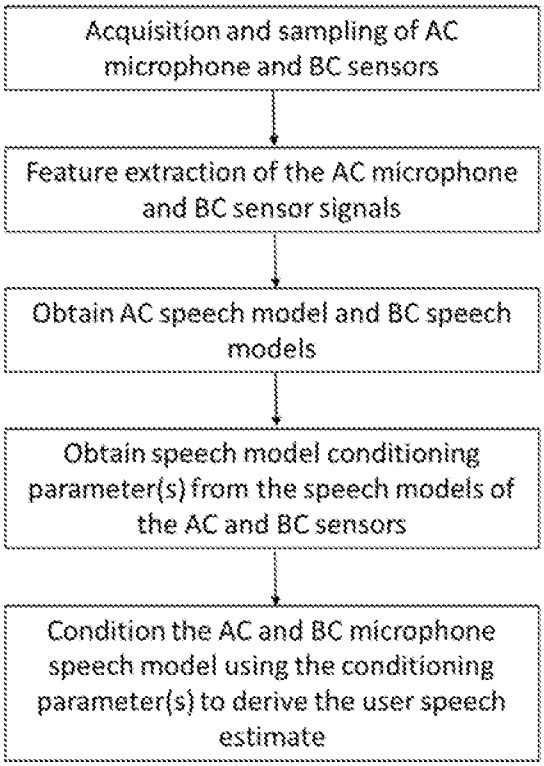
Acquisition and sampling of AC microphone and BC sensors
Feature extraction of the AC microphone and BC sensor signals
Obtain AC speech model and BC speech models
Obtain speech model conditioning parameter(s) from the speech models of the AC and BC sensors
Condition the AC and BC microphone speech model using the conditioning parameter(s) to derive the user speech estimate

Fig. 4

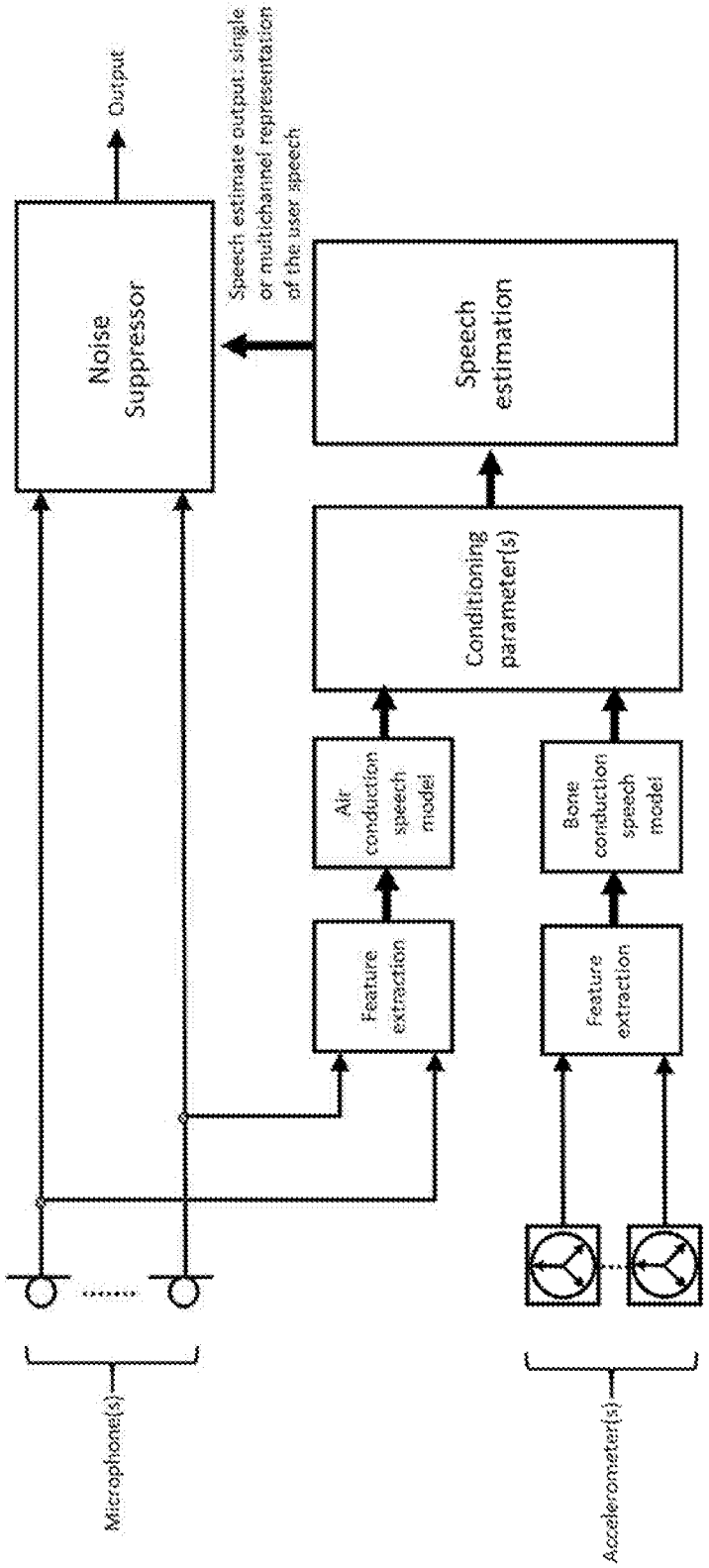
Output
Noise Suppressor
Speech estimate output: single or multichannel representation of the user speech
Speech estimation
Conditioning parameter(s)
Air conduction speech model
Bone conduction speech model
Feature extraction
Feature extraction
Microphone(s)
Accelerometer(s)

Fig. 5

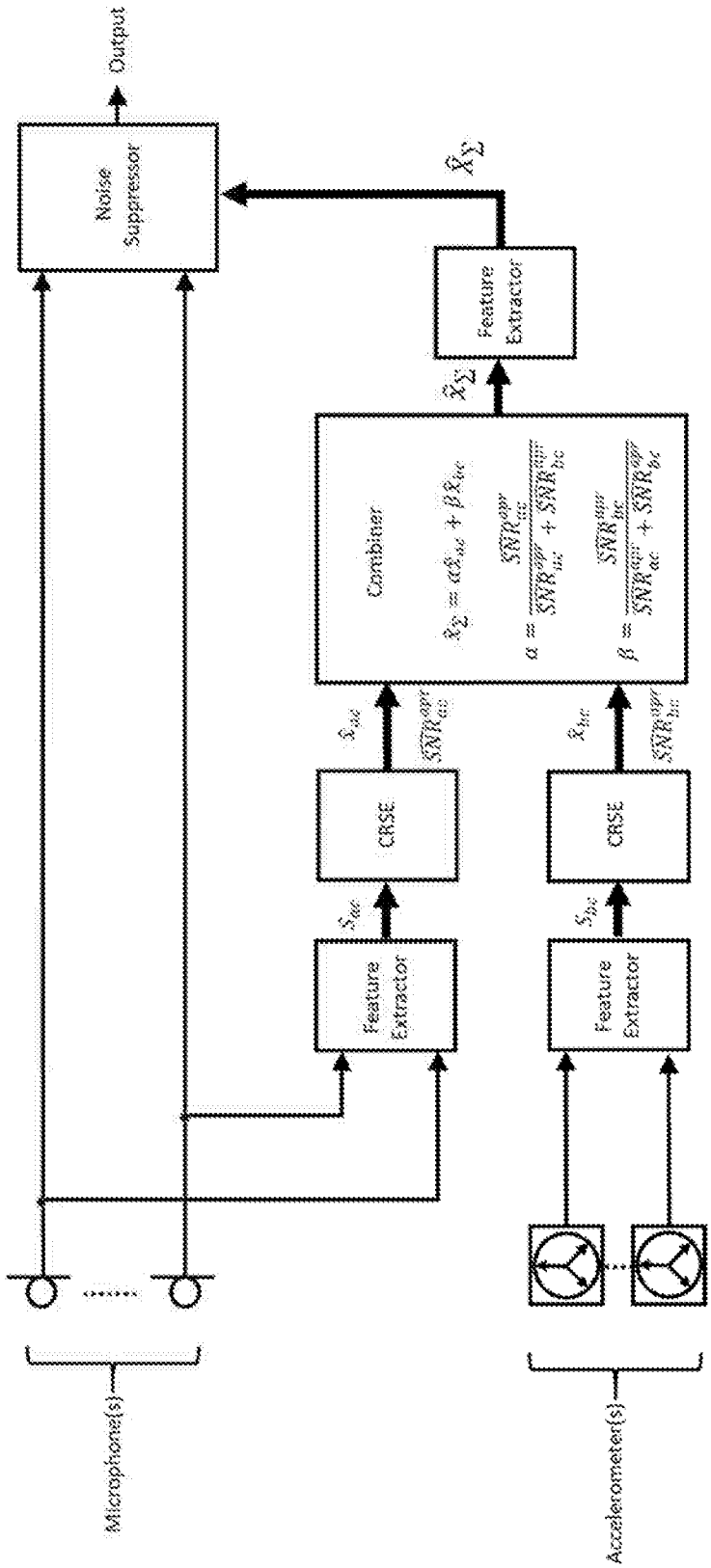
Output
Noise Suppressor
Feature Extractor
Combiner
CRSE
CRSE
Feature Extractor
Feature Extractor
Microphone(s)
Accelerometer(s)

Fig. 6

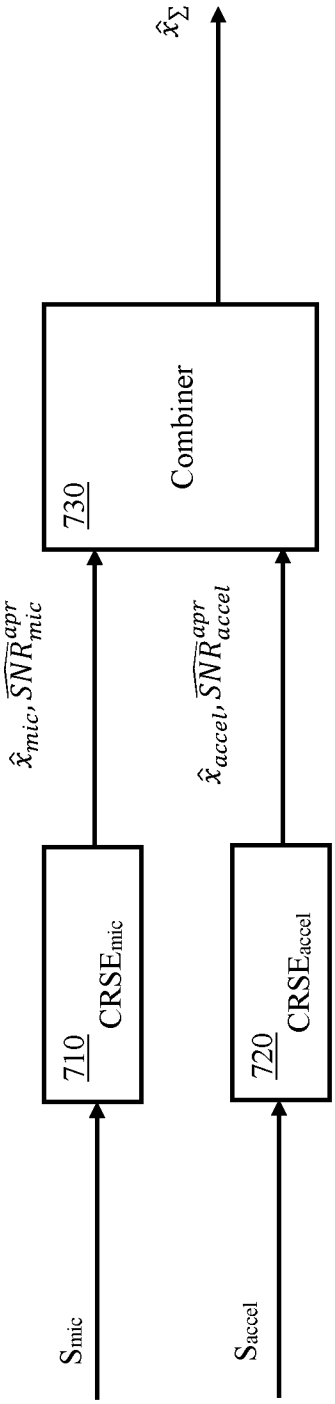
$\hat{x}_\Sigma$
730
Combiner
$\hat{x}_{mic}, \widehat{SNR}^{apr}_{mic}$
$\hat{x}_{accel}, \widehat{SNR}^{apr}_{accel}$
710
CRSEmic
720
CRSEaccel
Smic
Saccel

Fig. 7

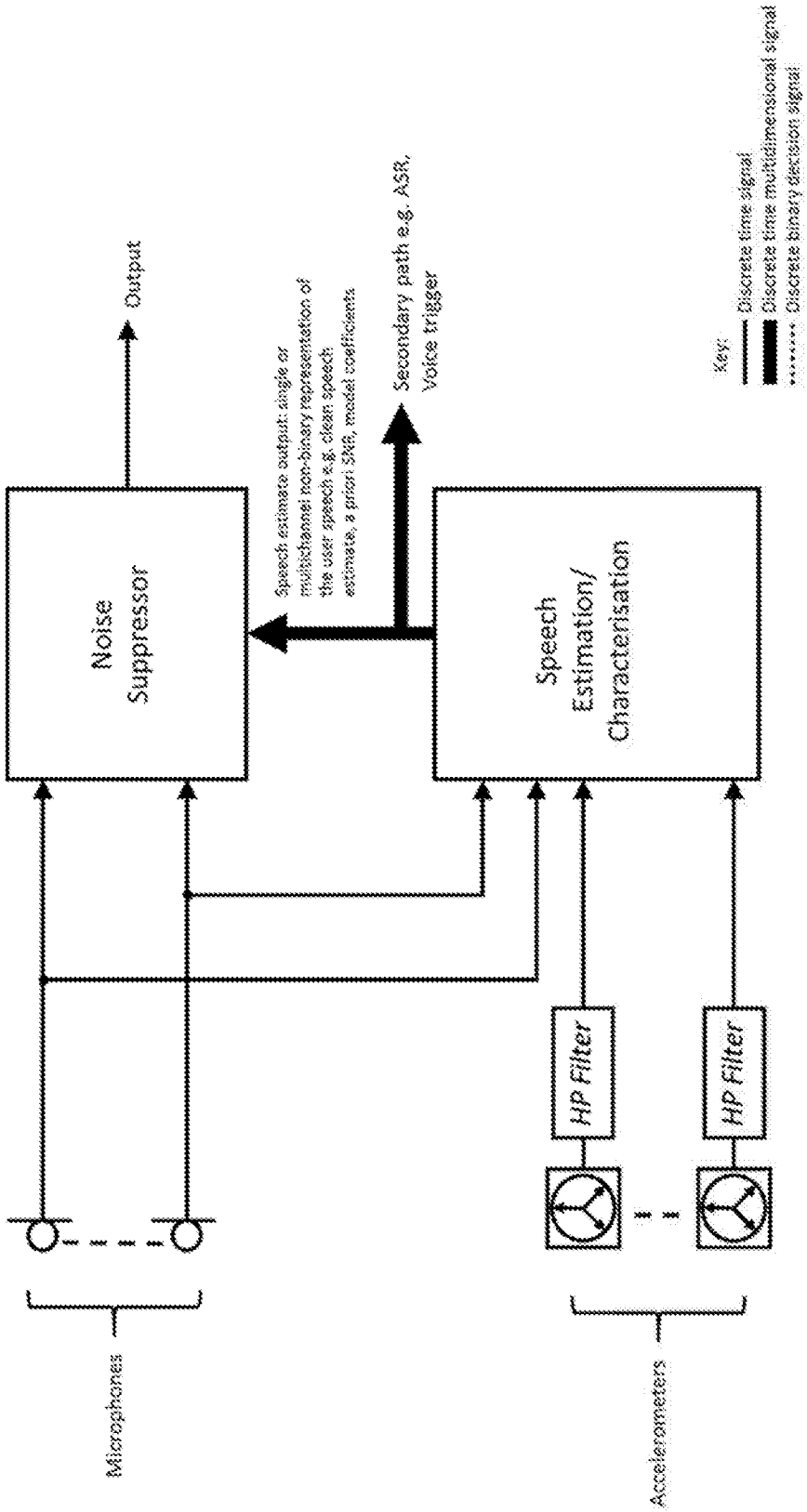
Output
Noise Suppressor
Speech estimate output: single or multichannel non-binary representation of the user speech e.g. clean speech estimate, a priori SNR, modal coefficients
Secondary path e.g. ASR, Voice trigger
Speech Estimation/ Characterisation
Key:
Discrete time signal
Discrete time multidimensional signal
Discrete binary decision signal
HP Filter
HP Filter
Microphones
Accelerometers

Fig. 8

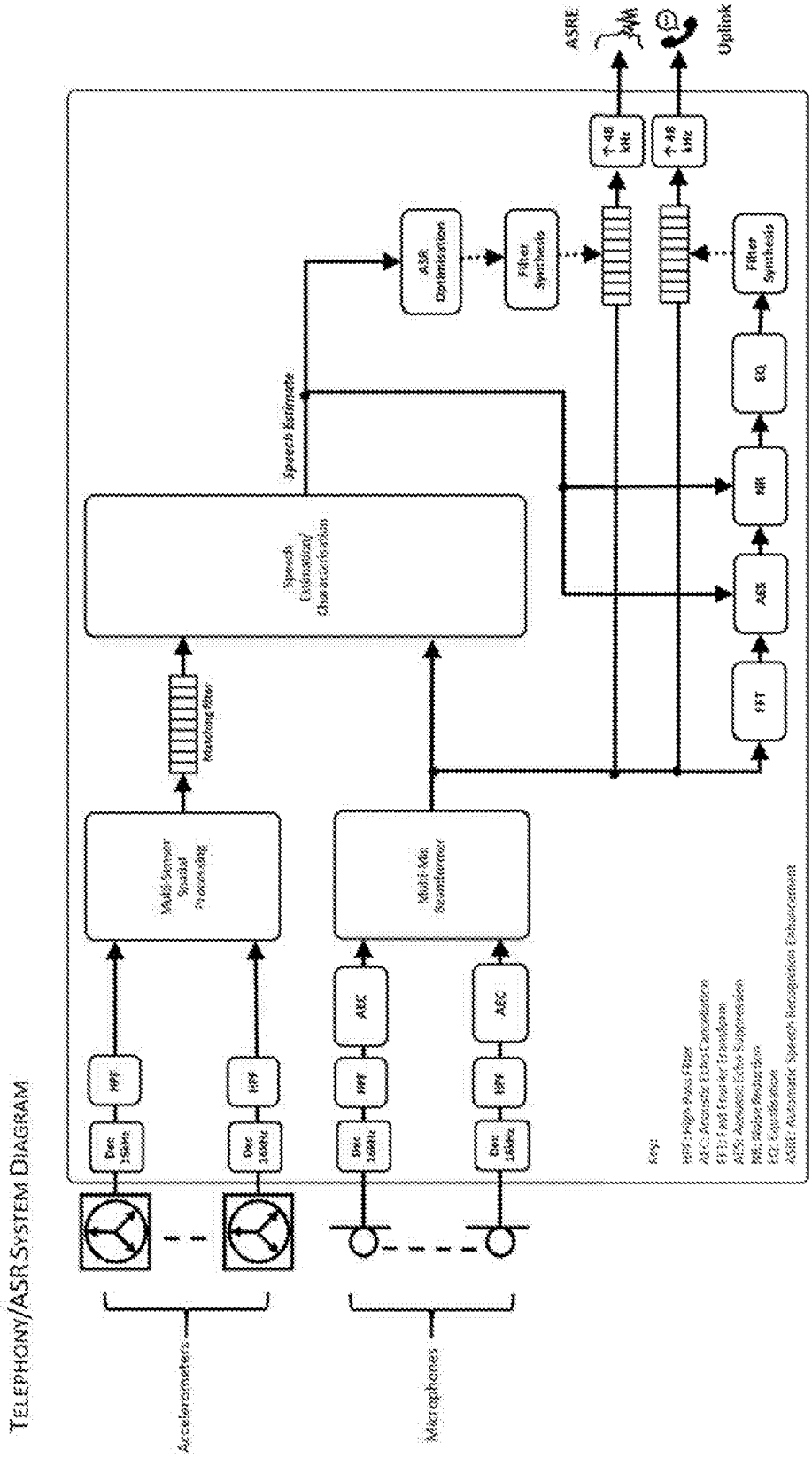
Telephony/ASR System Diagram
Accelerometers
Microphones
Multi-Sensor Spatial Processing
Multi-Mic Beamformer
Speech Estimate
Speech Enhancement/ Characterisation
ASR Optimisation
Filter Synthesis
ASRE
Uplink
FFT
AES
NR
EQ
HPF
AEC
Key:

Fig. 9

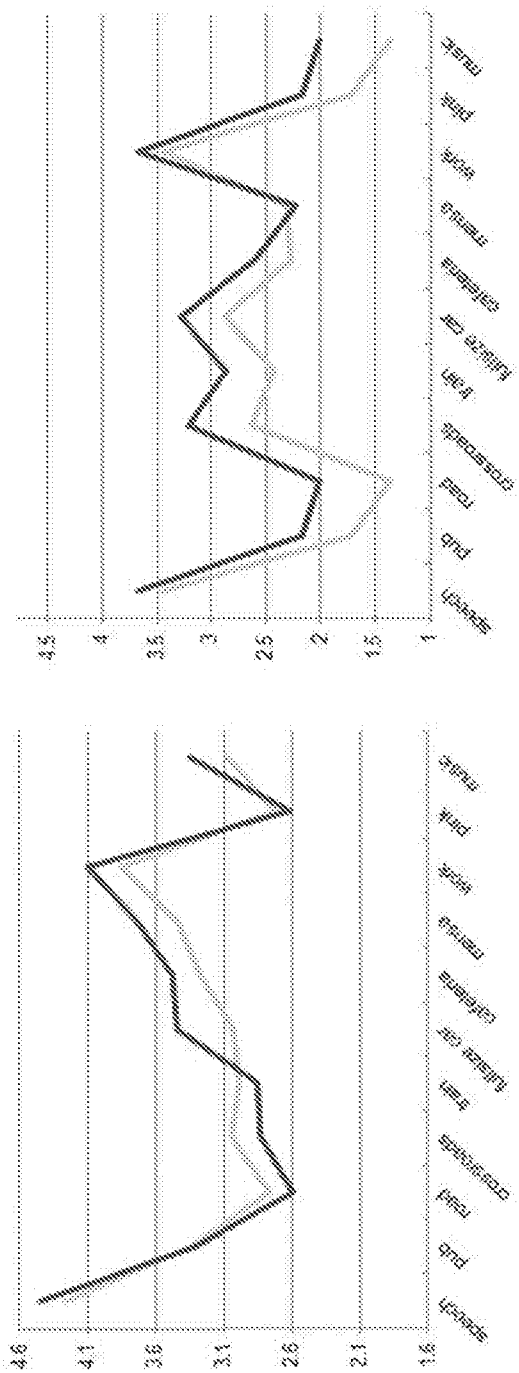
3.Quest S-MOS (WB)
4.6
4.1
3.6
3.1
2.6
2.1
1.6
3.Quest N-MOS (WB)
4.5
4
3.5
3
2.5
2
1.5
1
Noisy speech envelope without conditioning
Noisy speech envelope with conditioning

Fig. 10

INTERNATIONAL SEARCH REPORT

International application No
PCT/GB2018/051658

A. CLASSIFICATION OF SUBJECT MATTER
INV. H04R1/10 H04R1/46 G10L21/0208
ADD. H04R3/00 G10L21/0216

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)
H04R 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)
EPO-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	WO 2016/209530 A1 (INTEL IP CORP [US]) 29 December 2016 (2016-12-29)	1-5,10, 11, 16-27, 32-37, 42,43, 48-59, 64-66
Y	page 4, line 1 - page 5, line 25; figures 2-5	26, 28-31, 58,60-63
A		6-9, 12-15, 38-41, 44-47

-/--

[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
23 August 2018	31/08/2018

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	Navarri, Massimo

1

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/GB2018/051658

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	JP 2003 264883 A (DENSO CORP) 19 September 2003 (2003-09-19)	1,3-5, 10, 17-26, 32,33, 35-37, 42, 49-58, 64,65
Y	the whole document	27-29, 33, 59-61,65
A		6-9, 12-15, 38-41, 44-47

Y	US 2014/072148 A1 (SMITH WESLEY S [US] ET AL) 13 March 2014 (2014-03-13)	26-29, 33, 58-61,65
A	paragraphs [0012] - [0022]; figures 1-3	6-9, 12-15, 38-41, 44-47

Y	EP 2 811 485 A1 (FUJITSU LTD [JP]) 10 December 2014 (2014-12-10)	30,31, 62,63
A	paragraphs [0076] - [0086]; figures 17-18	6-9, 12-15, 38-41, 44-47

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/GB2018/051658

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2016209530 A1	29-12-2016	CN 107667401 A	06-02-2018
		EP 3314908 A1	02-05-2018
		JP 2018518696 A	12-07-2018
		KR 20180014187 A	07-02-2018
		TW 201712673 A	01-04-2017
		US 2016379661 A1	29-12-2016
		WO 2016209530 A1	29-12-2016
JP 2003264883 A	19-09-2003	NONE	
US 2014072148 A1	13-03-2014	CN 104604249 A	06-05-2015
		TW 201414325 A	01-04-2014
		US 2014072148 A1	13-03-2014
		WO 2014039243 A1	13-03-2014
EP 2811485 A1	10-12-2014	EP 2811485 A1	10-12-2014
		JP 6123503 B2	10-05-2017
		JP 2014239346 A	18-12-2014
		US 2014363020 A1	11-12-2014