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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
10 September 2004 (10.09.2004)

PCT

(10) International Publication Number
WO 2004/077090 A1

(51) International Patent Classification[7]: G01S 3/808

(21) International Application Number: PCT/DK2004/000077

(22) International Filing Date: 4 February 2004 (04.02.2004)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
PA 2003 00288 25 February 2003 (25.02.2003) DK

(71) Applicant *(for all designated States except US)*: **OTICON A/S** [DK/DK]; Strandvejen 58, DK-2900 Hellerup (DK).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **RASMUSSEN, Karsten, Bo** [DK/DK]; c/o OTICON A/S, Strandvejen 58, DK-2900 Hellerup (DK). **LAUGESEN, Søren** [DK/DK]; c/o Oticon A/S, Strandvejen 58, DK-2900 Hellerup (DK).

(74) Common Representative: **OTICON A/S**; Poulsen, Henning, Knak, Strandvejen 58, DK-2900 Hellerup (DK).

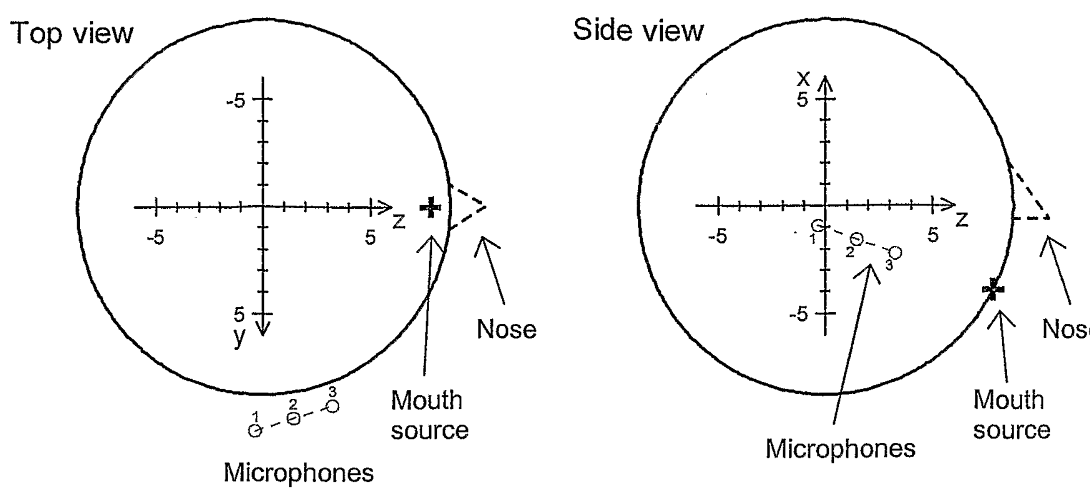
(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: METHOD FOR DETECTION OF OWN VOICE ACTIVITY IN A COMMUNICATION DEVICE



(57) Abstract: In the method according to the invention a signal processing unit receives signals from at least two microphones worn on the user's head, which are processed so as to distinguish as well as possible between the sound from the user's mouth and sounds originating from other sources. The distinction is based on the specific characteristics of the sound field produced by own voice, e.g. near-field effects (proximity, reactive intensity) or the symmetry of the mouth with respect to the user's head.

WO 2004/077090 A1

TITLE

Method for detection of own voice activity in a communication device.

AREA OF THE INVENTION

The invention concerns a method for detection of own voice activity to be used in connection with a communication device. According to the method at least two microphones are worn at the head and a signal processing unit is provided, which processes the signals so as to detect own voice activity.

The usefulness of own voice detection and the prior art in this field is described in DK patent application PA 2001 01461. This document also describes a number of different methods for detection of own voice.

However, it has not been proposed to base the detection of own voice on the sound field characteristics that arise from the fact that the mouth is located symmetrically with respect to the user's head. Neither has it been proposed to base the detection of own voice on a combination of a number individual detectors, each of which are error-prone, whereas the combined detector is robust.

BACKGROUND OF THE INVENTION

From DK PA 2001 01461 the use of own voice detection is known, as well as a number of methods for detecting own voice. These are either based on quantities that can be derived from a single microphone signal measured e.g. at one ear of the user, that is, overall level, pitch, spectral shape, spectral comparison of auto-correlation and auto-correlation of predictor coefficients, cepstral coefficients, prosodic features, modulation metrics; or based on input from a special transducer, which picks up vibrations in the ear canal caused by vocal activity. While the latter method of own voice detection is expected to be very reliable it requires a special transducer as described, which is expected to be difficult to realise. In contradiction, the former methods are readily implemented, but it has not been demonstrated or even theoretically substantiated that these methods will perform reliable own voice detection.

From US publication No.: US 2003/0027600 a microphone antenna array using voice activity detection is known. The document describes a noise reducing audio receiving system, which comprises a microphone array with a plurality of microphone elements for receiving an audio signal. An array filter is connected to the microphone array for filtering noise in accordance with select filter coefficients to develop an estimate of a speech signal. A voice activity detector is employed, but no considerations concerning far-field contra near-field are employed in the determination of voice activity.

From WO 02/098169 a method is known for detecting voiced and unvoiced speech using both acoustic and non-acoustic sensors. The detection is based upon amplitude differences between microphone signals due to the presence of a source close to the microphones.

The object of this invention is to provide a method, which performs reliable own voice detection, which is mainly based on the characteristics of the sound field produced by the user's own voice. Furthermore the invention regards obtaining reliable own voice detection by combining several individual detection schemes. The method for detection of own vice can advantageously be used in hearing aids, head sets or similar communication devices.

SUMMARY OF THE INVENTION

The invention provides a method for detection of own voice activity in a communication device wherein one or both of the following set of actions are performed,

A: providing at least two microphones at an ear of a person, receiving sound signals by the microphones and routing the signals to a signal processing unit wherein the following processing of the signal takes place: the characteristics, which are due to the fact that the microphones are in the acoustical near-field of the speaker's mouth and in the far-field of the other sources of sound are determined, and based on this characteristic it is assessed whether the sound signals originates from the users own voice or originates from another source,

B: providing at least a microphone at each ear of a person and receiving sound signals by the microphones and routing the microphone signals to a signal processing unit wherein the following processing of the signals takes place: the characteristics, which are due to the fact that the user's mouth is placed symmetrically with respect to the user's head are determined, and based on this characteristic it is assessed whether the sound signals originates from the users own voice or originates from another source.

The microphones may be either omni-directional or directional. According to the suggested method the signal processing unit in this way will act on the microphone signals so as to distinguish as well as possible between the sound from the user's mouth and sounds originating from other sources.

In a further embodiment of the method the overall signal level in the microphone signals is determined in the signal processing unit, and this characteristic is used in the assessment of whether the signal is from the users own voice. In this way knowledge of normal level of speech sounds is utilized. The usual level of the users voice is recorded, and if the signal level in a situation is much higher or much lower it is than taken as an indication that the signal is not coming from the users own voice.

According to an embodiment of the method, the characteristics, which are due to the fact that the microphones are in the acoustical near-field of the speaker's mouth are determined by a filtering process in the form of FIR filters, the filter coefficients of which are determined so as to maximize the difference in sensitivity towards sound coming from the mouth as opposed to sound coming from all directions by using a Mouth-to-Random-far-field index (abbreviated $M2R$) whereby the M2R obtained using only one microphone in each communication device is compared with the M2R using more than one microphone in each hearing aid in order to take into account the different source strengths pertaining to the different acoustic sources. This method takes advantage of the acoustic near field close to the mouth.

In a further embodiment of the method the characteristics, which are due to the fact that the user's mouth is placed symmetrically with respect to the user's head are determined by receiving the signals $x_1(n)$ and $x_2(n)$, from microphones positioned at each ear of the user, and compute the cross-correlation function between the two signals: $R_{x_1x_2}(k) = E\{x_1(n)x_2(n-k)\}$, applying a detection criterion to the output $R_{x_1x_2}(k)$, such that if the maximum value of $R_{x_1x_2}(k)$ is found at $k = 0$ the dominating sound source is in the median plane of the user's head whereas if the maximum value of $R_{x_1x_2}(k)$ is found elsewhere the dominating sound source is away from the median plane of the user's head. The proposed embodiment utilizes the similarities of the signals received by the hearing aid microphones on the two sides of the head when the sound source is the users own voice.

The combined detector then detects own voice as being active when each of the individual characteristics of the signal are in respective ranges.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a schematic representation of a set of microphones of an own voice detection device according to the invention.

Figure 2 is a schematic representation of the signal processing structure to be used with the microphones of an own voice detection device according to the invention.

Figure 3 shows in two conditions illustrations of metric suitable for an own voice detection device according to the invention.

Figure 4 is a schematic representation of an embodiment of an own voice detection device according to the invention.

Figure 5 is a schematic representation of a preferred embodiment of an own voice detection device according to the invention.

DESCRIPTION OF PREFERRED EMBODIMENTS

Figure 1 shows an arrangement of three microphones positioned at the right-hand ear of a head, which is modelled as a sphere. The nose indicated in Figure 1 is not part of the model but is useful for orientation. Figure 2 shows the signal processing structure to be used with the three microphones in order to implement the own voice detector. Each microphone signal as digitised and sent through a digital filter (W_1, W_2, W_3), which may be a FIR filter with L coefficients. In that case, the summed output signal in Figure 2 can be expressed as

$$y(n) = \sum_{m=1}^{M} \sum_{l=0}^{L-1} w_{ml} x_m (n-l) = \underline{w}^{\mathrm{T}} \underline{x},$$

where the vector notation

$$\underline{w} = \left[w_{10} \cdots w_{ML-1} \right]^{\mathrm{T}}, \ \underline{x} = \left[x_1(n) \cdots x_M(n-L+1) \right]^{\mathrm{T}}$$

has been introduced. Here M denotes the number of microphones (presently $M = 3$) and w_{ml} denotes the l th coefficient of the m th FIR filter. The filter coefficients in $\underline{w}$ should be determined so as to distinguish as well as possible between the sound from the user's mouth and sounds originating from other sources. Quantitatively, this is accomplished by means of a metric denoted $\Delta M2R$, which is established as follows. First, Mouth-to-Random-far-field index (abbreviated $M2R$) is introduced. This quantity may be written as

$$M2R(f) = 10 \log_{10} \left(\frac{\left| Y_{Mo}(f) \right|^2}{\left| Y_{Rff}(f) \right|^2} \right),$$

where $Y_{Mo}(f)$ is the spectrum of the output signal $y(n)$ due to the mouth alone, $Y_{Rff}(f)$ is the spectrum of the output signal $y(n)$ averaged across a representative set of far-field sources and f denotes frequency. Note that the $M2R$ is a function of frequency and is given in dB. The $M2R$ has an undesirable dependency on the source strengths of both the far-field and mouth sources. In order to remove this dependency a reference $M2R_{ref}$ is introduced, which is the $M2R$ found with the front microphone alone. Thus the actual metric becomes

$$\Delta M2R(f) = M2R(f) - M2R_{ref}(f).$$

Note that the ratio is calculated as a subtraction since all quantities are in dB, and that it is assumed that the two component $M2R$ functions are determined with the same set of far-field and mouth sources. Each of the spectra of the output signal $y(n)$, which goes into the calculation of $\Delta M2R$, can be expressed as

$$Y(f) = \sum_{m=1}^{M} W_m(f) Z_{Sm}(f) q_S(f),$$

where $W_m(f)$ is the frequency response of the m th FIR filter, $Z_{Sm}(f)$ is the transfer impedance from the sound source in question to the m th microphone and $q_S(f)$ is the source strength. Thus, the determination of the filter coefficients $\underline{w}$ can be formulated as the optimisation problem

$$\max_{\underline{w}} |\Delta M2R|,$$

where $|\cdot|$ indicates an average across frequency. The determination of $\underline{w}$ and the computation of $\Delta M2R$ has been carried out in a simulation, where the required transfer impedances corresponding to Figure 1 have been calculated according to a spherical head model. Furthermore, the same set of filters have been evaluated on a set of transfer impedances measured on a Brüel & Kjær HATS manikin equipped with a prototype set of microphones. Both set of results are shown in the left-hand side of Figure 3. In this figure a $\Delta M2R$-value of 0 dB would indicate that distinction between sound from the mouth and sound from other far-field sources was impossible, whereas positive values of $\Delta M2R$ indicates possibility for distinction. Thus, the simulated result in Figure 3(left) is very encouraging. However, the result found with measured transfer impedances is far below the simulated result at low frequencies. This is because the optimisation problem so far has disregarded the issue of robustness. Hence, robustness is now taken into account in terms of the White Noise Gain of the digital filters, which is computed as

$$WNG(f) = 10\log_{10}\left(\sum_{m=1}^{M} \left|W_m(e^{-j2\pi f/f_S})\right|^2\right),$$

where f_S is the sampling frequency. By limiting WNG to be within 15 dB the simulated performance is somewhat reduced, but much improved agreement is obtained between simulation and results from measurements, as is seen from the right-hand side of Figure 3. The final stage of the preferred embodiment regards the application of a detection criterion to the output signal $y(n)$, which takes place in the Detection block shown in Figure 2. Alternatives to the above $\Delta M2R$-metric are obvious, e.g. metrics based on estimated components of active and reactive sound intensity.

Considering an own voice detection device according to the invention, Figure 4 shows an arrangement of two microphones, positioned at each ear of the user, and a signal processing structure which computes the cross-correlation function between the two signals $x_1(n)$ and $x_2(n)$, that is,

$$R_{x_1x_2}(k) = E\{x_1(n)x_2(n-k)\}.$$

As above, the final stage regards the application of a detection criterion to the output $R_{x_1x_2}(k)$, which takes place in the Detection block shown in Figure 4. Basically, if the maximum value of $R_{x_1x_2}(k)$ is found at $k=0$ the dominating sound source is in the median plane of the user's head and may thus be own voice, whereas if the maximum value of $R_{x_1x_2}(k)$ is found elsewhere the dominating sound source is away from the median plane of the user's head and cannot be own voice.

Figure 5 shows an own voice detection device, which uses a combination of individual own voice detectors. The first individual detector is the near-field detector as described above, and as sketched in Figure 1 and Figure 2. The second individual detector is based on the spectral shape of the input signal $x_3(n)$ and the third individual detector is based on the overall level of the input signal $x_3(n)$. In this example the combined own voice detector is thought to flag activity of own voice when all three individual detectors flag own voice activity. Other combinations of individual own voice detectors, based on the above described examples, are obviously possible. Similarly, more advanced ways of combining the outputs from the individual own voice detectors into the combined detector, e.g. based on probabilistic functions, are obvious.

CLAIMS

1. Method for detection of own voice activity in a communication device whereby one or both of the following set of actions are performed,
    - A: providing at least two microphones at an ear of a person, receiving sound signals by the microphones and routing the signals to a signal processing unit wherein the following processing of the signal takes place: the characteristics, which are due to the fact that the microphones are in the acoustical near-field of the speaker's mouth and in the far-field of the other sources of sound are determined, and based on this characteristic it is assessed whether the sound signals originates from the users own voice or originates from another source,
    - B: providing at least a microphone at each ear of a person and receiving sound signals by the microphones and routing the microphone signals to a signal processing unit wherein the following processing of the signals takes place: the characteristics, which are due to the fact that the user's mouth is placed symmetrically with respect to the user's head are determined, and based on this characteristic it is assessed whether the sound signals originates from the users own voice or originates from another source.

2. Method as claimed in claim 1, whereby the overall signal level in the microphone signals is determined in the signal processing unit, and this characteristic is used in the assessment of whether the signal is from the users own voice.

3. Method as claimed in claim 1, whereby the characteristics, which are due to the fact that the microphones are in the acoustical near-field of the speaker's mouth are determined by a filtering process in the form of FIR filters, the filter coefficients of which are determined so as to maximize the difference in sensitivity towards sound coming from the mouth as opposed to sound coming from all directions by using a Mouth-to-Random-far-field index (abbreviated $M2R$) whereby the M2R obtained using only one microphone in each hearing aid is compared with the M2R using

more than one microphone in each hearing aid in order to take into account the different source strengths pertaining to the different acoustic sources.

4. Method as claimed in claim 4 wherein M2R is determined in the following way:

$$M2R(f) = 10\log_{10}\left(\frac{\left|Y_{Mo}(f)\right|^2}{\left|Y_{Rff}(f)\right|^2}\right),$$

where $Y_{Mo}(f)$ is the spectrum of the output signal $y(n)$ due to the mouth alone, $Y_{Rff}(f)$ is the spectrum of the output signal $y(n)$ averaged across a representative set of far-field sources and f denotes frequency.

5. Method as claimed in claim 1, whereby the characteristics, which are due to the fact that the user's mouth is placed symmetrically with respect to the user's head are determined by receiving the signals $x_1(n)$ and $x_2(n)$, from microphones positioned at each ear of the user, and compute the cross-correlation function between the two signals: $R_{x_1x_2}(k) = E\{x_1(n)x_2(n-k)\}$, applying a detection criterion to the output $R_{x_1x_2}(k)$, such that if the maximum value of $R_{x_1x_2}(k)$ is found at $k = 0$ the dominating sound source is in the median plane of the user's head whereas if the maximum value of $R_{x_1x_2}(k)$ is found elsewhere the dominating sound source is away from the median plane of the user's head.

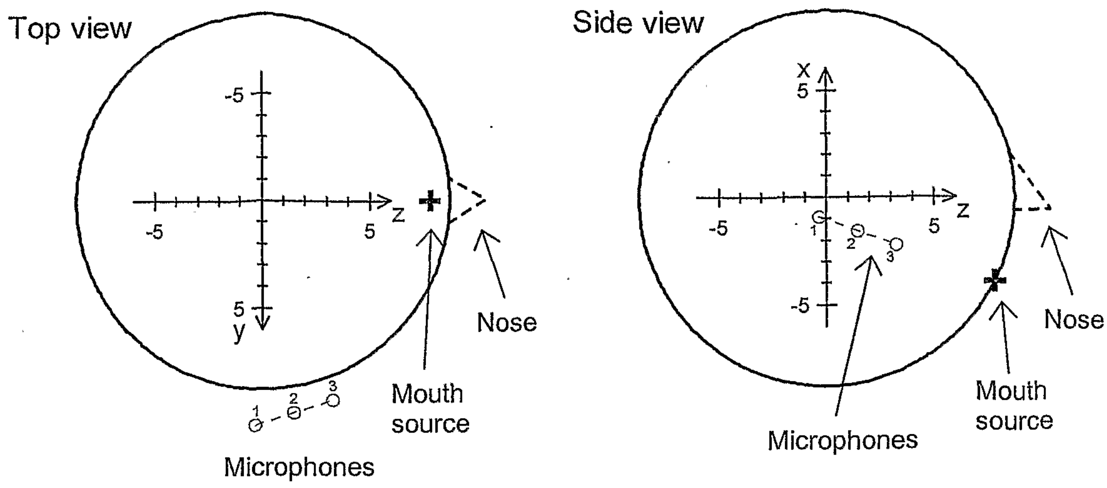


Figure 1.

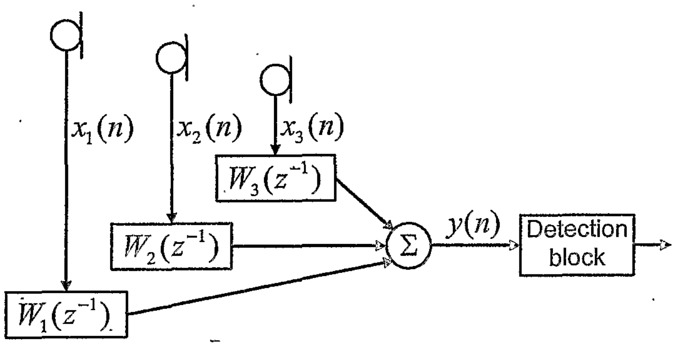


Figure 2.

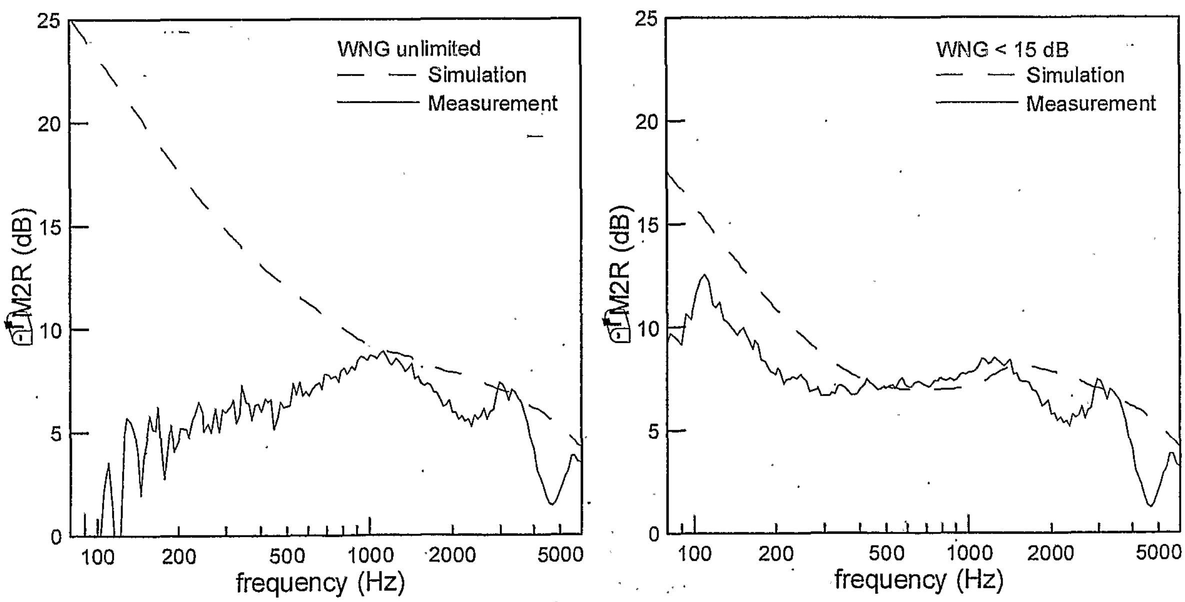


Figure 3.

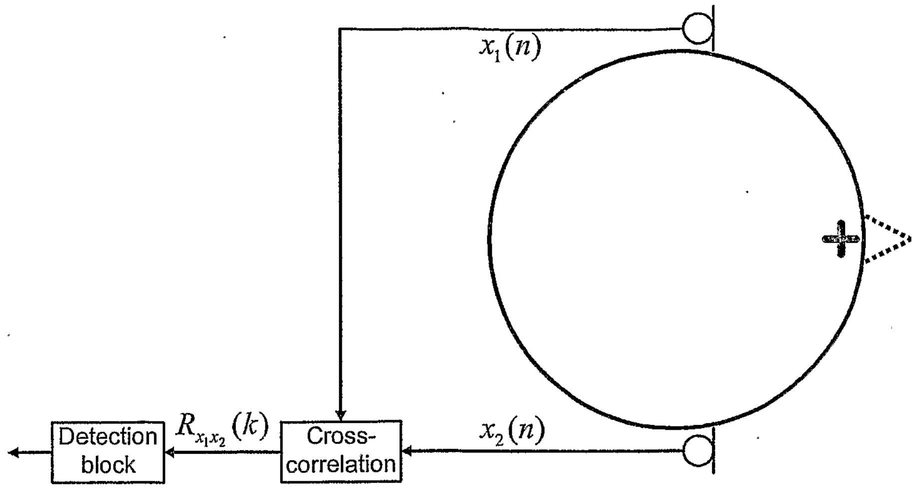


Figure 4.

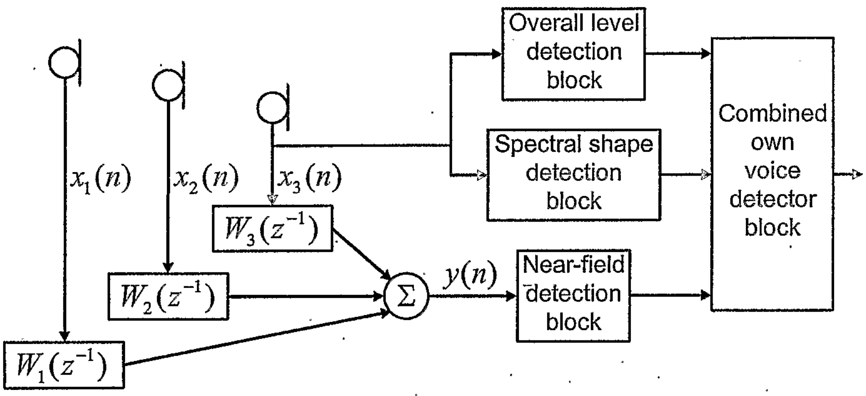


Figure 5.

INTERNATIONAL SEARCH REPORT

International Application No
PCT/DK2004/000077

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G01S3/808

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)
IPC 7 G01S 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 practical, search terms used)

EPO-Internal, WPI Data, PAJ, INSPEC

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
P,X	LAUGESEN S ET AL: "Design of a microphone array for headsets" IEEE WORKSHOP ON APPLICATIONS OF SIGNALS PROCESSING TO AUDIO AND ACOUSTICS, 19 October 2003 (2003-10-19), pages 37-40, XP010696436 NEW PALTZ, NY the whole document ----- -/--	1-4

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

[X] Patent family members are listed in 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 document 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
24 May 2004	14/06/2004

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, Tx. 31 651 epo nl, Fax: (+31–70) 340–3016	Santos Luque, R

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5 448 637 A (YAMAGUCHI ET AL) 5 September 1995 (1995-09-05) abstract column 2, line 32 - line 36 column 3, line 24 - line 39 column 5, line 25 - line 29 column 6, line 40 - line 48 column 7, line 1 - line 51 column 12, line 66 - column 13, line 6 figures 4-8	1-4
X	NORDHOLM S E ET AL: "Chebyshev optimization for the design of broadband beamformers in the near field" IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II: ANALOG AND DIGITAL SIGNAL PROCESSING, JAN. 1998, IEEE, USA, vol. 45, no. 1, January 1998 (1998-01), pages 141-143, XP002281203 ISSN: 1057-7130 the whole document	1-4
X	THOMAS M. SULLIVAN: "Multi-Microphone Correlation-Based Processing for Robust Automatic Speech Recognition" PH.D THESIS, CARNEGIE MELLON UNIVERSITY, August 1996 (1996-08), XP002281204 PITTSBURG, PENNSYLVANIA page 19, line 4 - page 20, line 5	1,5
X	US 2001/019516 A1 (WAKE ET AL) 6 September 2001 (2001-09-06) abstract page 1, paragraph 2 - paragraph 5 figures 1-3	1,5
A	RYAN J G ET AL: "Array optimization applied in the near field of a microphone array" IEEE TRANS. SPEECH AUDIO PROCESS. (USA), IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, MARCH 2000, IEEE, USA, vol. 8, no. 2, March 2000 (2000-03), pages 173-176, XP002281205 ISSN: 1063-6676 page 173	1-5

-/--

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	KNAPP C H ET AL: "The generalized correlation method for estimation of time delay" IEEE TRANSACTIONS ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, USA, vol. ASSP-24, no. 4, August 1976 (1976-08), pages 320-327, XP002281206 ISSN: 0096-3518 abstract page 320, right-hand column, line 17 -----	1,5
A	US 2003/027600 A1 (KRASNY ET AL) 6 February 2003 (2003-02-06) cited in the application abstract -----	1-5

INTERNATIONAL SEARCH REPORT

Information on patent family members

Int tional Application No
PCT/DK2004/000077

Patent document cited in search report		Publication date	Patent family member(s)		Publication date
US 5448637	A	05-09-1995	JP	2769554 B2	25-06-1998
			JP	6189388 A	08-07-1994
			US	5606607 A	25-02-1997
			US	5664014 A	02-09-1997
US 2001019516	A1	06-09-2001	JP	2001236092 A	31-08-2001
US 2003027600	A1	06-02-2003	NONE		