(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 2 312 579 A1

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
20.04.2011 Bulletin 2011/16

(51) Int Cl.:
G10L 21/02 *(2006.01)*

(21) Application number: **09173163.8**

(22) Date of filing: **15.10.2009**

(84) Designated Contracting States:
AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO SE SI SK SM TR
Designated Extension States:
AL BA RS

(71) Applicant: **Honda Research Institute Europe GmbH**
63073 Offenbach/Main (DE)

(72) Inventor: **Heckmann, Martin**
60316, Frankfurt (DE)

(74) Representative: **Rupp, Christian**
Mitscherlich & Partner
Patent- und Rechtsanwälte
Sonnenstrasse 33
80331 München (DE)

(54) Speech from noise separation with reference information

(57) System and method for separating a mixture signal containing a mixture of target information and interfering information comprising means (2) for receiving the mixture signal, means (3) for receiving a reference signal and a signal processing unit (1) configured to extract cues from the reference signal and to separate the target information from the mixture signal using the cues.

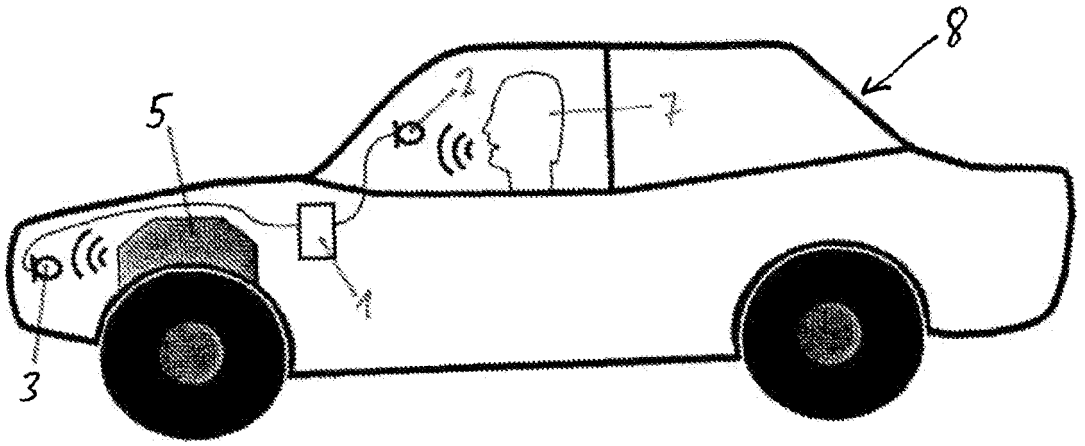

Fig. 2

EP 2 312 579 A1

Description

[0001] The invention generally refers to the processing of acoustically sensed signals.

[0002] The present invention relates to a system and a method for separating a mixture signal containing a mixture of acoustical target information ("speech") and interfering information ("noise").

[0003] In many everyday situations different background noise sources are present while we talk on the phone or try to operate a device via speech. This noise, however, makes speech recognition more difficult for humans and especially for machines. For instance, headset free telecommunication in a car or the operation of different devices in the car (e.g. navigation systems, radio) via speech are often interfered by the driving noise (e.g. engine noise) of the car. A similar situation is found when riding a motorcycle, because the noise generated by the motorcycle engine significantly impairs the quality of telecommunication during a ride. Another area where speech recognition is more and more used is robotics. Background noise (e.g. caused by a fan to cool the robot's CPU) is here also reducing the quality of speech recognition.

[0004] Therefore, different approaches have already been proposed to reduce the noise and thereby improve the speech signal. The European patent application EP 1 879 180 A1 for example shows a method to reduce the background noise in speech signals with the help of a reference microphone. The main idea behind this method is to estimate the spectrum of the noise based on a reference microphone which captures only the noise and then subtract these spectral components from the microphone signal which captures the mixture of the speech and noise signal. The main disadvantage, however, of this method is that the acoustic environment where the noise is captured is normally different from that where the mixture of the noise and the speech signal are captured (e.g. the engine compartment and the passenger compartment if one wants to reduce the engine noise in a car). As a consequence not the noise at it was present in the passenger compartment is subtracted from the corresponding mixture signal but the noise as it was recorded in the engine compartment. To remedy this, a filter is used which replicates the transformation of the signal it underwent on its way from the engine compartment to the passenger compartment.

[0005] However, this filtering operation is highly dependent on the position of the speaker and has hence to be adaptive. This, on the other hand, is problematic as no error signal is available to adapt the filter when the speaker is talking.

[0006] There also exist approaches to enhance speech signals in a way inspired by the processing in the human brain. They are commonly referred to as Computational Auditory Scene Analysis (CASA). For humans it was observed that they are able to separate different concurrent sound sources and focus on one source. The underlying mechanisms are able to separate signals based on cues which can serve to bind or group different time frequency regions to one acoustic source. Such cues are e.g. fundamental frequency, location in space, or common on- and off-sets.

[0007] Different systems have already been developed which are able to separate a speech signal from an interfering signal based on such cues, e.g. fundamental frequency or common on- and off-sets. For doing so these cues are estimated from the signal containing the mixture of the different sources present. This step usually comprises the split of this mixture signal into different frequency channels. Based on the mentioned cues it is determined next for each of the frequency channels at each instant in time to which of the detected sources the signal belongs. The results obtained are usually quite good but also entail significant degradations of the speech signal. One reason for this is that the cues used for separating the sound sources (e.g. fundamental frequency, on- and off-sets) have to be extracted from the mixture of the sources and hence this extraction process is error prone. Similar the number of present sources also has to be estimated from the mixture signal.

[0008] It is therefore the object of the present invention to propose a system and a method to improve noise reduction in a signal that contains a mixture of a noise and a speech signal.

[0009] This object is achieved by means of the features of the independent claims. The dependent claims develop further the central idea of the present invention.

[0010] The present invention relates to a technique for reducing noise in a mixture signal containing a mixture of noise and speech by means of additional reference information captures e.g. by a second microphone. In contrast to previous methods the present invention, however, does not try to reduce the noise directly in the signal domain but uses techniques inspired by Computational Auditory Scene Analysis (CASA) to reduce the noise.

[0011] Therefore, a system for the separation of a mixture signal containing a mixture of target information and interfering information is proposed, that comprises means for receiving the mixture signal, means for receiving a reference signal and a signal processing unit configured to extract cues from the reference signal and to separate the target information from the mixture signal using these cues.

[0012] In addition, a method for separating a mixture signal containing a mixture of target information and interfering information is proposed, said comprises the steps of receiving the mixture signal, receiving a reference signal and extracting cues from the reference signal and separating the target information from the mixture signal using these cues.

[0013] The means for receiving the mixture signal and the means for receiving the reference signal may comprise a microphone and a recording unit each, wherein the microphone for the mixture signal may be positioned close to the origin of the target information and the mi-

crophone for the reference signal may be positioned close the origin of the interfering information, when the means for receiving the reference signal are configured to receive interfering information. The interfering information can be also extracted from the speed of an engine.

[0014] Furthermore the means for receiving the reference signal may be configured to receive target information, wherein this information can be extracted from a video signal, in particular from the movement of a speaker's body or the speaker's lip movement in the video signal.

[0015] The signal processing unit of the system may comprise means for splitting the reference signal and the mixture signal in a multitude of frequency channels, means for extracting grouping cues from the reference signal and evaluating the grouping cues in the mixture signal for each frequency channel at each instant in time and means for allocating each frequency channel of the mixture signal at each instant in time to either the target information or the interfering information and separating the mixture signal into the target information and the interfering information.

[0016] In another embodiment the signal processing unit of the system comprises means for splitting the mixture signal in a multitude of frequency channels, means for extracting grouping cues from the reference signal and evaluating the grouping cues in the mixture signal at each instant in time and means for allocating each frequency channel of the mixture signal at each instant in time to either the target information or the interfering information and separating the mixture signal into the target information and the interfering information.

[0017] The grouping cues may be the fundamental frequency or on- or off-sets. The target information may be speech and the interfering information may be noise.

[0018] The system for separating a mixture signal may be included in a motorcycle helmet, wherein the means for receiving the mixture signal are positioned inside the helmet and the means for receiving the reference signal are positioned partly inside the helmet and partly close to the engine of a motorcycle, wherein the means for receiving the reference signal are connected via a cable or wireless.

[0019] These and other aspects and advantages of the present invention will become more apparent when studying the following detailed description, in connection with the figures, in which

Fig. 1 shows a motorcyclist with a helmet that includes a system according to the invention driving a motorcycle;

Fig. 2 shows a car with a driver including a system according to the invention;

Fig. 3 shows a method according to the invention.

[0020] Fig. 1 shows a motorcyclist 7 driving a motorcycle 6 and wearing a helmet 4. The helmet 4 includes a system according to the invention. The system comprises a signal processing unit 1, means for receiving a mixture signal, here a microphone 2, and means for receiving a reference signal, here a microphone 3a, and a receiving unit 3b. The microphone 2 for receiving the mixture signal and the receiving unit 3b are connected to the signal processing unit 1 via a cable. The microphone 3a for receiving the reference signal is, however, not positioned in the helmet 4, but close to the engine 5 of the motorcycle 6 to be at the origin of the interfering signal, which in the shown example may be the harmonic noise generated by the engine 5. The transmission of the reference signal of the microphone 3a to the receiving unit 3b can be for example accomplished via a wireless transmission.

[0021] The microphone 2 for receiving the mixture signal is positioned to the front of the helmet 4 close to the mouth of the motorcyclist 7. The microphone 2 is therefore positioned close to the origin of the target signal, here the acoustically sensed speech signal of the motorcyclist 7. However, the microphone 2 also receives noise of the engine 5, due to the fact that the engine 5 of the motorcycle 6 is quite loud and the engine noise is only slightly attenuated by the helmet 4. Therefore the mixture signal received by the microphone 2 contains a mixture of speech and noise.

[0022] The signal processing unit 1 is configured to extract cues from the reference signal received by the microphone 3 and to separate the speech from the mixture signal received by the microphone 2 using the cues. A detailed description of the extraction and separation will be given in combination with the method and Fig. 3.

[0023] The system according to the invention is therefore able to significantly reduce the engine noise in the mixture signal. As a result of the reduction telecommunication while riding will be improved.

[0024] Another application area for the system according to the invention is shown in Fig. 2, where a car 8 is shown including a similar system to that in Fig. 1. However, the signal processing unit 1 and the microphone 2 for receiving the mixture signal are not positioned in a helmet, but inside the car 8. In particular the microphone 2 for receiving the mixture signal is positioned in the passenger compartment to be near to the mouth of the driver 7. The microphone 3a for receiving the reference signal is again positioned close to the engine 5. The signal processing unit 1 can be positioned anywhere in the car and has connections to the microphone 2 for receiving the mixture signal and the microphone 3a for receiving the reference signal. Therefore a receiving unit 3b is not needed here. The system according to the invention does not only improve the headset free telecommunication but also speech based operation of devices in a car. In particular the reduction of the harmonic noise generated by the engine is here helpful.

[0025] Fig. 3 shows a method according to an embodiment of the invention. At the beginning an acoustically

sensed mixture signal and a reference signal are received (100, 101).

[0026] The reference signal is preferably directly sensed from the origin of the noise (e.g. an engine, fan, ...). In an ideal setup only the reference signal without any additional signals would be sensed such that the reference signal is available without distortions. This can best be achieved by sensing the reference signal close to its source.

[0027] In a same way the target signal will also preferably be sensed close to its source. In the case of a speech signal of the driver of a car or a motorcycle sensing close to the drivers mouth would be best. However, to allow a speech interaction where the driver does not need to wear any special device, i.e. headset free, the target signal is commonly sensed at a certain distance from its source. As a consequence only a mixture of the target signal and other sound sources is sensed. Hence, in one application of the present invention this mixture signal would be a mixture of noise generated by the engine and a speech signal of the driver. As already described above, the mixture signal and reference signal can for example be sensed by microphones.

[0028] After receiving the reference signal and the mixture signal both signals are split into a multitude of adjacent frequency channels 102, 103. For each frequency channel at each instant in time grouping auditory scene analysis cues are extracted from the reference signal ("noise signal") 105. These cues which are typically used in Computational Auditory Scene Analysis (CASA) systems for the separation of sources can be e.g. one or more of:

- fundamental frequency
- common on- and off-sets
- common modulation / fate
- spatial cues (ITD/ILD i.e. perceived origin)
- continuity
- sequential similarity.

[0029] These auditory cues provide information on the reference signal. Knowing these cues allows identifying the reference signal in the mixture signal. For doing so, these cues are extracted in the reference signal, where the reference signal is mostly undistorted and these cues can easily be extracted, and then evaluated in the mixture signal. As a result of this evaluation parts, i.e. frequency channels at each instant in time, can be identified in which the reference signal is dominating the mixture signal.

[0030] For the extraction of these auditory cues the reference signal is transformed into the frequency domain and they are determined for each frequency channel at each instance in time. For some of these cues as e.g. the fundamental frequency it is also possible to extract these cues directly in the time domain and then calculate their effect on the different frequency channels (in the case of the fundamental frequency signal parts will be present at the fundamental frequency and at its harmonics which can easily be calculated from the fundamental frequency). After extracting the auditory cues from the reference signal they are evaluated in the mixture signal (comprising e.g. noise and speech). After transforming the mixture signal into the frequency domain and the cues are evaluated for each frequency channel at each instant in time (104). Then it is possible to allocate each frequency channel at each instant in time to either the speech or the noise (106). Based on this allocation the mixture signal is separated, in discrete time steps, into frequency channels "speech" and frequency channels "noise" (107).

[0031] With this method it is now possible to measure the Computational Auditory Scene Analysis (CASA) cues from an undistorted noise signal and can then use this information to eliminate the noise in the mixture of speech and noise without the need to estimate the transfer function between the site of the recording of the noise and the mixture signal (e. g. from the engine compartment to the passenger compartment).

[0032] In Figs. 1 and 2 the system according to the invention is included in a motorcycle helmet and in a car. However, it is also possible that such a system is for example included in a robot. This would help to improve speech recognition systems in robots. Therefore robots or any other technical systems which are controlled by speech or interpret speech can be even used in loud and noisy environments.

[0033] Another application area where the system according to the invention can be used is the field of hearing devices. The elimination of a noise in a mixture signal that the hearing device is receiving helps the person who uses the hearing device to even better understand the speech of other persons.

[0034] The examples in Figs. 1 and 2 are showing a system where the reference signal uses noise from an engine. However it is also possible that for the reference signal information on the speech signal can be obtained e. g. by using a bone conductive microphone. In this case the necessary grouping information is extracted from the speech signal and then used to separate the speech signal from the noise signal.

Prior Art References

[0035]

[1] H. Puder, F. Steffens. Improved Noise Reduction for HandsFree Car Phones Utilizing Information on Vehicle and Engine Speeds. EUSIPCO, 2000

[2] EP1879180 - Reduction of background noise in hands- free Systems

[3] Bregman, A. Auditory Scene Analysis MIT Press, 1990

[4] Brown, G. J. & Cooke, M. P. Computational Auditory Scene Analysis Computer Speech and Language, 1994, 1, 297-336

[5] Heckmann, M.; Joublin, F. & Körner, E. Sound

Source Separation for a Robot Based on Pitch Proc IEEE/RSJ Int . 1 Conf. on Robots and Intell. Syst., 2005, 203- 208

[6] Hu, G. & Wang, D. L. Monaural Speech Segregation Based on Pitch Tracking and Amplitude Modulation IEEE Trans. Neural Networks, 2004, 15, 1135-1150

[7] Hu, G. & Wang, D. Auditory segmentation based on onset and offset analysis IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15, 396- 405

Claims

1. System for the separation of an acoustical target signal from an acoustically sensed mixture signal containing a mixture of said target signal and interfering signal, such as noise, the system comprising

- means (2) for acoustically sensing the mixture signal,
- means (3) for sensing a reference signal, preferably close to a known source of noise, , and
- a signal processing unit (1) configured to
  - extract cues from the reference signal for each instance in time at for each of a plurality of preferably adjacent and continuous frequency channels, and
  - to separate the target signal from the mixture signal for each of said frequency channels and each instance in time, using the cues.

2. System according to claim 1, **characterized in that,** the means for receiving the mixture signal comprising a microphone (2), wherein the microphone (2) is positioned close to the origin of the target information.

3. System according to any of claims 1 to 2, **characterized in that,** the means (3) for receiving the reference signal are configured to receive interfering information.

4. System according to claim 3, **characterized in that,** the means (3) for receiving the reference signal comprising a microphone (3a), wherein the microphone (3a) is positioned close to the origin of the interfering information.

5. System according to any of claims 1 to 2, **characterized in that,** the means (3) for receiving the reference signal are configured to receive target information.

6. System according to any of claims 1 to 5, **characterized in that,** the signal processing unit (1) comprises

- means for splitting the reference signal and the mixture signal in a multitude of frequency channels,
- means for extracting grouping cues from the reference signal and evaluating the grouping cues in the mixture signal for each frequency channel at each instant in time and
- means for allocating each frequency channel of the mixture signal at each instant in time to either the target information or the interfering information and separating the mixture signal into the target information and the interfering information.

7. System according to any of claims 1 to 5, **characterized in that,**

the signal processing unit (1) comprises

- means for splitting the mixture signal in a multitude of frequency channels,
- means for extracting grouping cues from the reference signal and evaluating the grouping cues in the mixture signal at each instant in time and
- means for allocating each frequency channel of the mixture signal at each instant in time to either the target information or the interfering information and separating the mixture signal into the target information and the interfering information.

8. A robot, an air/land/sea vehicle, a voice-controlled system or an artificial hearing aid, comprising a system according to any of the preceding claims.

9. A method for separating a mixture signal containing a mixture of target information and interfering information, comprising the steps

- Receiving the mixture signal (100),
- Receiving a reference signal (101) and
- Extracting cues from the reference signal and separating the target information from the mixture signal using the cues (102-107).

10. The method according to claim 9, **characterized in that,** the target information is speech and the interfering information is noise.

11. Method according to any of claims 9 to 10, **characterized in that,** the step of extracting cues from the reference signal

and separating the target information from the mixture signal using the cues comprises the steps

- Splitting the mixture signal in a multitude of frequency channels (102),
- Splitting the reference signal in a multitude of frequency channels (103),
- Extracting grouping cues from the reference signal for each frequency channel at each instant in time (105),
- Evaluating grouping cues in the mixture signal for each frequency channel at each instant in time (104),
- Allocating each frequency channel at each instant in time to either the target information or the interfering information based on the evaluation of the grouping cues (106) and
- Separating the mixture of the target information and the interfering information based on the previous allocation (107).

12. Method according to any of claims 9 or 10, **characterized in that,** the step of extracting cues from the reference signal and separating the target information from the mixture signal using the cues comprises the steps

- Splitting the mixture signal in a multitude of frequency channels,
- Extracting grouping cues from the reference signal at each instant in time,
- Evaluating grouping cues in the mixture signal for each frequency channel at each instant in time ,
- Allocating each frequency channel at each instant in time to either the target information or the interfering information based on the evaluation of the grouping cues and
- Separating the mixture of the target information and the interfering information based on the previous allocation.

13. The method according to any of claims 11 to 12, **characterized in that,** the fundamental frequency is used as grouping cue.

14. The method according to any of claims 11 to 12, **characterized in that,** on- or off-sets are used as grouping cue.

15. The method according to any of claims 9 to 14, **characterized in that,** the reference signal contains interfering information.

16. The method according to claim 15, **characterized in that,** the interfering information of the reference signal is extracted from the speed of an engine (5) .

17. The method according to any of claims 9 to 14, **characterized in that,** the reference signal contains target information.

18. The method according to claim 17, **characterized in that,** the target information of the reference signal is extracted from the movements of a speaker's body or lip movements of a speaker in a video signal.

19. A computer software program product, performing a method according to any of claims 9 to 18 when run on a computing unit.

20. A motorcycle helmet (4), being provided with a system according to any of the claims 1 to 7.

21. A motorcycle helmet according to claim 20, **characterized in that,** the means (2) for receiving the mixture signal are positioned inside the helmet (4) and the means (3) for receiving the reference signal are positioned partly (3b) inside the helmet (4) and partly (3a) close to a engine (5) of a motorcycle (6), wherein the means (3a, 3b) for receiving the reference signal are connected via a cable or wireless.

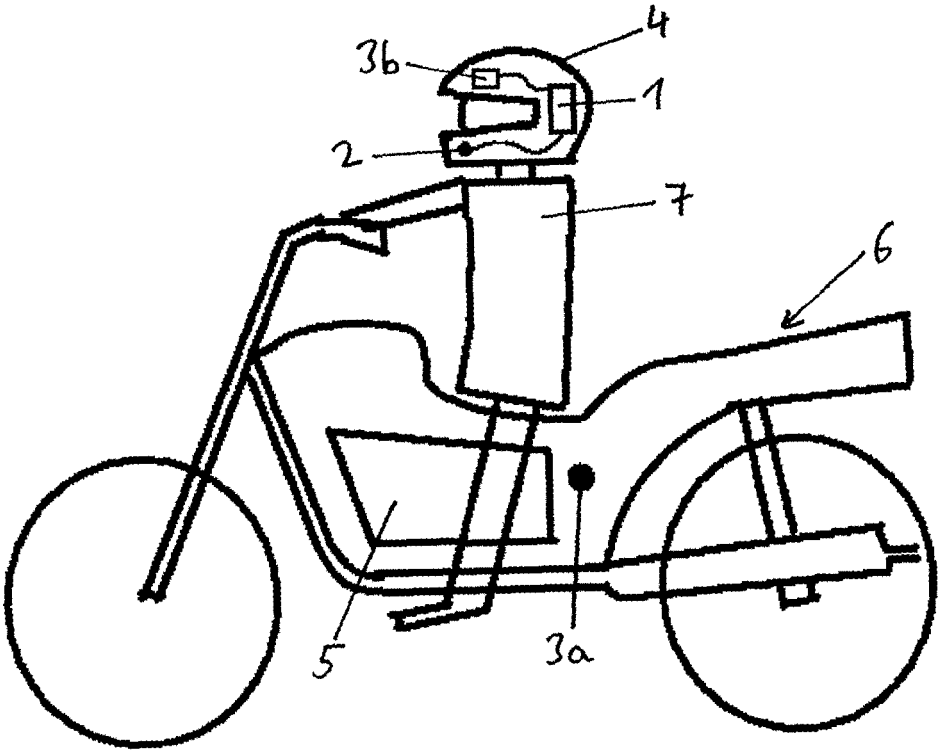

Fig. 1

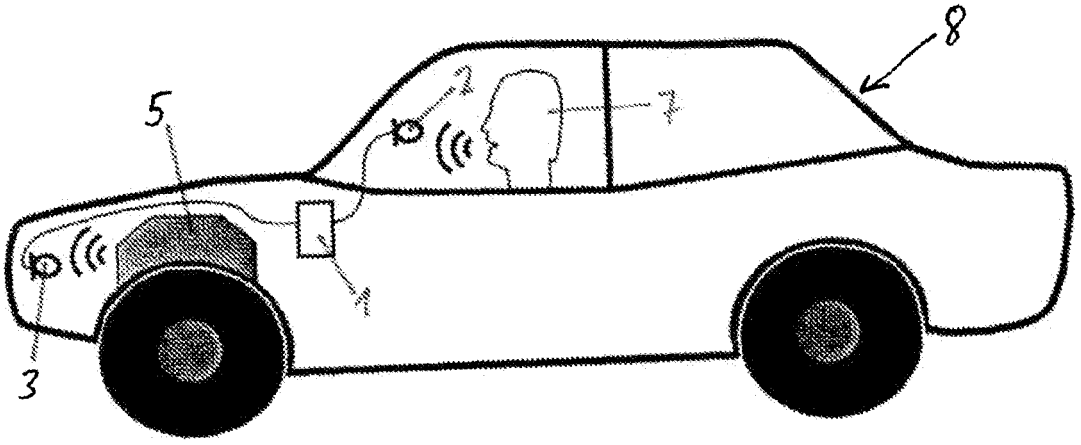

Fig. 2

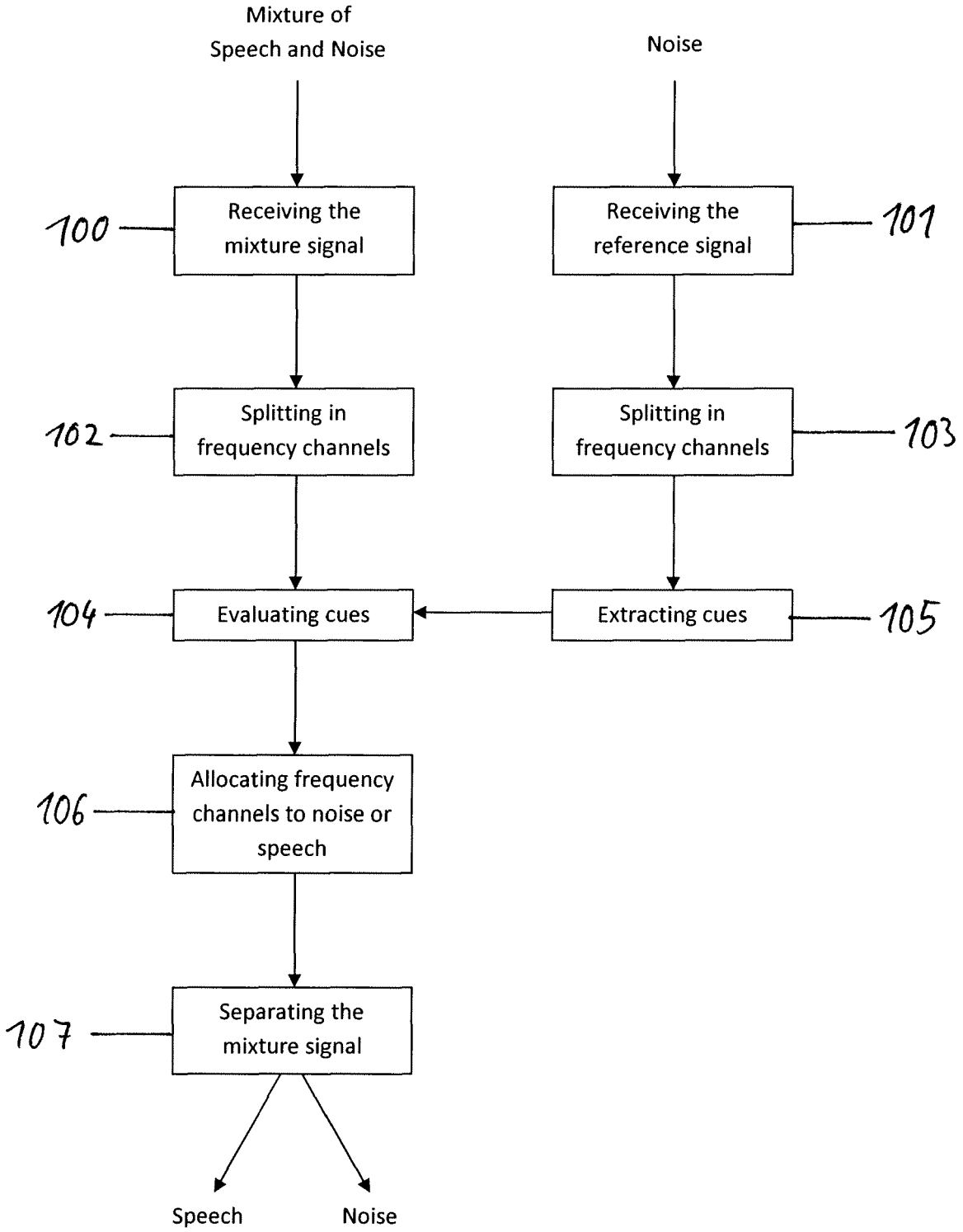
Mixture of
Speech and Noise
Noise
100
Receiving the mixture signal
Receiving the reference signal
101
102
Splitting in frequency channels
Splitting in frequency channels
103
104
Evaluating cues
Extracting cues
105
106
Allocating frequency channels to noise or speech
107
Separating the mixture signal
Speech
Noise

Fig. 3

Europäisches Patentamt
European Patent Office
Office européen des brevets

EUROPEAN SEARCH REPORT

Application Number
EP 09 17 3163

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (IPC)
Y	WO 98/01956 A2 (CHIEFS VOICE INC [US]) 15 January 1998 (1998-01-15) * claim 1 * * page 12, lines 2-9 *	1-21	INV. G10L21/02
Y	US 4 932 063 A (NAKAMURA SHOGO [JP]) 5 June 1990 (1990-06-05) * column 1, line 62 - column 3, line 9 *	1-21	
A	KUO S M ET AL: "ACTIVE NOISE CONTROL: A TUTORIAL REVIEW" PROCEEDINGS OF THE IEEE, IEEE. NEW YORK, US, vol. 87, no. 6, 1 June 1999 (1999-06-01), pages 943-973, XP011044219 ISSN: 0018-9219 *Section III*	1-21	
A,D	PUDER H ET AL: "Improved noise reduction for hands-free car phones utilizing information on vehicle and engine speeds" SIGNAL PROCESSING : THEORIES AND APPLICATIONS, PROCEEDINGS OFEUSIPCO, XX, XX, vol. 3, 1 January 2000 (2000-01-01), pages 1851-1854, XP009030255 *Section 3.1*	1-21	TECHNICAL FIELDS SEARCHED (IPC) G10L
A	GB 2 377 805 A (20 20 SPEECH LTD [GB]) 22 January 2003 (2003-01-22) * page 18, lines 7-14 *	18	

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
The Hague	10 March 2010	Bensa, Julien

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)

ANNEX TO THE EUROPEAN SEARCH REPORT ON EUROPEAN PATENT APPLICATION NO. EP 09 17 3163

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report. The members are as contained in the European Patent Office EDP file on
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

10-03-2010

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 9801956 A2	15-01-1998	AU 719596 B2	11-05-2000
		AU 3895497 A	02-02-1998
		CA 2259641 A1	15-01-1998
		EP 0901726 A2	17-03-1999
		JP 2000514618 T	31-10-2000
		US 6072881 A	06-06-2000
US 4932063 A	05-06-1990	DE 3837066 A1	11-05-1989
		GB 2212035 A	12-07-1989
		JP 1118900 A	11-05-1989
GB 2377805 A	22-01-2003	NONE	

EPO FORM P0459

For more details about this annex : see Official Journal of the European Patent Office, No. 12/82

REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Patent documents cited in the description

- EP 1879180 A1 **[0004]**
- EP 1879180 A **[0035]**

Non-patent literature cited in the description

- **Bregman, A.** *Auditory Scene Analysis MIT Press,* 1990 **[0035]**
- **Brown, G. J ; Cooke, M. P.** *Computational Auditory Scene Analysis Computer Speech and Language,* 1994, vol. 1, 297-336 **[0035]**
- **Heckmann, M. ; Joublin, F. ; Körner, E.** Sound Source Separation for a Robot Based on Pitch Proc IEEE/RSJ Int . 1 Conf. *Robots and Intell. Syst.,* 2005, 203-208 **[0035]**
- **Hu, G. ; Wang, D. L.** Monaural Speech Segregation Based on Pitch Tracking and Amplitude Modulation. *IEEE Trans. Neural Networks,* 2004, vol. 15, 1135-1150 **[0035]**
- **Hu, G. ; Wang, D.** *Auditory segmentation based on onset and offset analysis IEEE Transactions on Audio, Speech, and Language Processing,* 2007, vol. 15, 396-405 **[0035]**