



(51) Classification internationale des brevets :

G10L 21/0364 (2013.01) G10L 25/78 (2013.01)
H04R 1/10 (2006.01) H04R 3/00 (2006.01)

(21) Numéro de la demande internationale :

PCT/EP2017/061931

(22) Date de dépôt international :

18 mai 2017 (18.05.2017)

(25) Langue de dépôt :

français

(26) Langue de publication :

français

(30) Données relatives à la priorité :

1655019 02 juin 2016 (02.06.2016) FR

(71) Déposant : PARROT DRONES [FR/FR] ; 174-178 quai
de Jemmapes, 75010 PARIS (FR).

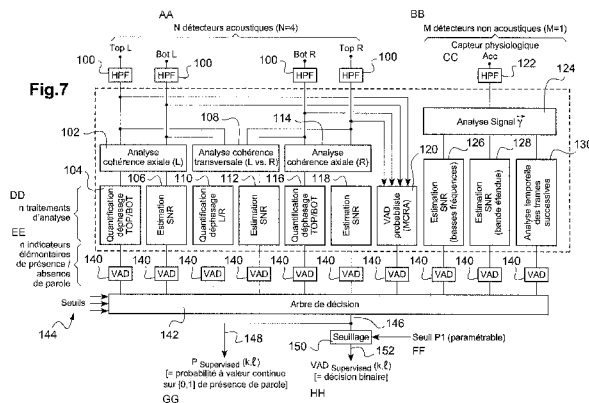
(72) Inventeurs : VITTE, Guillaume ; 64 rue Amelot, 75011 Paris (FR). REMOND, Fabien ; 78 Avenue Larroumes, 94240 L'Hay-Les-Roses (FR). RENARD, Nicolas ; 35 Rue de la Bienfaisance, Esc C, 2ème étage, 75008 Paris (FR). BRIOT, Alexandre ; 7 Passage des marais, 75010 Paris (FR).

(74) Mandataire : DUPUIS-LATOURE, Dominique et al. ; 10 Boulevard Haussmann, 75009 PARIS (FR).

(81) États désignés (sauf indication contraire, pour tout titre de protection nationale disponible) : 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, 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,

(54) Title: AUDIO MICROPHONE/HEADSET COMBINATION COMPRISING MULTIPLE MEANS FOR DETECTING VOCAL ACTIVITY WITH SUPERVISED CLASSIFIER

(54) Titre : COMBINE AUDIO MICRO/CASQUE COMPRENANT DES MOYENS DE DETECTION D'ACTIVITE VOCALE MULTIPLES A CLASSIFIEUR SUPERVISE



AA N acoustic detectors (N=4)
BB M non-acoustic detectors (M=1)
CC Physiological sensor
DD n analysis processes
EE n elementary indicators of presence/absence of speech
FF Threshold P1 (configurable)
GG P_{supervised}(k,l) [probability of continuous value on [0,1] of presence of speech]
HH VAD_{supervised}(k,l) [binary decision]
124 Signal analysis
102 Axial coherence analysis (L)
108 Transverse coherence analysis (L vs. R)
114 Axial coherence analysis (R)
104 Quantification of TOP/BOT phase difference
106 SNR estimation
110 Quantification of L/R phase difference
120 VAD probabilistic (MCRA)
126 Estimation of SNR (low frequencies)
128 SNR estimation (extended band)
130 Temporal analysis of successive frames
142 Decision tree
144 Thresholds
150 Thresholding

(57) Abstract: The headset supports a network of acoustic and non-acoustic detectors issuing respective signals (TopL, BotL, TopR, BotR, Acc) comprising a noise component and, if applicable, a component of speech of the wearer of the headset. A plurality of n distinct analysis processes (104, 106, 110, 112, 116, 118, 120, 126, 128, 130) for detecting vocal activity are carried out on the detected signals and/or combinations of these signals, processes which issue, at the output, a corresponding plurality of n elementary indicators quantifying respective different values of continuous speech presence probability. Means for classifying the observation vector formed by the n elementary indicators comparing these n elementary indicators with a set of successive predefined thresholds (144) classified according to a decision tree Boolean configuration (142), and generating in response a global indicator of the presence or absence of speech (P_{supervised}(k,l); VAD_{supervised}(k,l)).

(57) Abrégé : Le casque supporte un réseau de détecteurs acoustiques et non-acoustiques délivrant des signaux respectifs (TopL, BotL, TopR, BotR, Acc) comportant une composante de bruit et le cas échéant une composante de parole du porteur du casque. Une pluralité de n traitements d'analyse distincts (104, 106, 110, 112, 116, 118, 120, 126, 128, 130) de détection d'activité vocale sont effectués sur les signaux captés et/ou des combinaisons de ces signaux, traitements qui délivrent en sortie une pluralité correspondante de n indicateurs élémentaires quantifiant des valeurs différentes respectives de probabilité à valeur continue de présence de parole. Des moyens classifieurs du vecteur d'observation formé par les n indicateurs élémentaires comparent ces n

PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

- (84) **États désignés** (*sauf indication contraire, pour tout titre de protection régionale disponible*) : ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), eurasién (AM, AZ, BY, KG, KZ, RU, TJ, TM), européen (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).

Publiée:

— avec rapport de recherche internationale (Art. 21(3))

**Combiné audio micro/casque comprenant des moyens
de détection d'activité vocale multiples à classifieur supervisé**

L'invention concerne un ensemble de type "micro-casque", comprenant un casque audio, combiné avec un ensemble microphonique apte à capter la voix du porteur du casque.

Le casque comprend généralement deux écouteurs réunis par un arceau, 5 chaque écouteur comprenant une coque fermée logeant un transducteur de reproduction sonore. Cette coque entoure l'oreille (casque "circum-aural") ou est en appui sur celle-ci (casque "supra-aural"), avec interposition d'un coussinet isolant. Il existe également des écouteurs de type dit "intra-auriculaire" ou "intra-conque" avec un élément à placer dans le 10 conduit auditif et débordant dans le creux du pavillon de l'oreille au-delà du conduit auditif, sans coussinet entourant ou recouvrant l'oreille. Dans la suite, on fera indifféremment référence à des écouteurs réunis par un arceau, où le transducteur est logé dans une coque entourant l'oreille (casque "circum-aural") ou en appui sur celle-ci (casque "supra-aural") ou 15 avec des écouteurs de type "intra-auriculaire" ou analogue.

Dans tous les cas, le casque peut être utilisé pour l'écoute d'une source audio (musique par exemple) provenant d'un appareil tel qu'un lecteur MP3, radio, *smartphone*, etc. auquel il est relié par une liaison filaire ou bien par une liaison sans fil, notamment de type *Bluetooth*. Grâce à l'en- 20 semble microphonique, il est également possible, en complément de l'écoute de la source audio d'utiliser ce casque pour des fonctions de communication telles que des fonctions de téléphonie "mains libres" et/ou de fonctions de reconnaissance vocale. Le transducteur du casque reproduit alors la voix du locuteur distant avec lequel le porteur du casque (locuteur proche) est en conversation. 25

Un tel ensemble combiné micro-casque est par exemple décrit dans les EP 2 518 724 A1, EP 2 930 942 A1 et EP 2 945 399 A1 (tous trois au nom de Parrot).

Lorsque le microphone est utilisé par le porteur du casque, il convient que 30 le niveau du bruit d'environnement (d'origine acoustique) présent sur le signal capté par l'ensemble microphonique soit le plus faible possible, de manière à préserver la qualité subjective de la voix du porteur du casque

(locuteur proche) sur ce signal, afin de transmettre au locuteur distant un signal fidèle et intelligible au travers de la communication téléphonique.

Le paramètre de référence pour évaluer cet impératif est le rapport signal sur bruit, ci-après "SNR", qui est le rapport du signal utile (correspondant à la voix du porteur du casque), sur le bruit qui l'entoure, d'origine essentiellement acoustique.

Il s'agit d'un paramètre important pour évaluer notamment l'efficacité d'un algorithme de débruitage appliqué au signal de voix du porteur du casque (locuteur proche). Un tel algorithme est mis en œuvre pour réduire le bruit d'environnement, d'origine acoustique, présent sur ce signal, mais doit aussi préserver la qualité subjective de la voix du porteur du casque, de façon à transmettre au locuteur distant un signal fidèle et intelligible dans toutes les conditions de bruit auxquelles le porteur du casque pourra être exposé.

En premier lieu, lors d'une conversation téléphonique, le réflexe consistant à parler plus fort dans un environnement bruyant (effet Lombard) agit déjà sur le niveau de la voix et sur le contenu fréquentiel de la voix (élévation de la hauteur et aplatissement du spectre) en fonction du bruit perçu.

De plus, divers effets psycho-acoustiques et de positionnement des capteurs microphoniques ont une influence notable sur le niveau du SNR. À cet égard, plusieurs configurations peuvent ainsi être utilement comparées.

La configuration la plus favorable est celle d'un "kit filaire", où le microphone est monté sur un fil reliant le casque au *smartphone* du porteur du casque. Ce placement est avantageux pour la captation de la voix car le microphone est alors positionné juste au-dessous du menton, donc près de la bouche, avec au surplus la possibilité pour le porteur du casque de rapprocher le microphone de sa bouche dans une situation bruyante.

Cette configuration est comparable, du point de vue du SNR, à celle d'un téléphone porté contre la joue (*smartphone* avec microphone intégré).

Pour une situation typique de communication téléphonique nomade, en contexte urbain, on observe généralement un SNR de l'ordre de +3 dB avec un téléphone porté contre la joue ou un kit filaire.

Par comparaison, la configuration d'un combiné micro-écouteur de type "oreillette *Bluetooth*", où le microphone est alors proche de l'oreille, présente une dégradation du SNR, de l'ordre de -3 dB.

5 Une configuration encore plus défavorable est celle d'un casque sans fil fermé de type circumaural, où le microphone est situé sur l'oreillette et le casque isole les deux oreilles du porteur du casque. Le SNR est alors typiquement de l'ordre de -7 à -9 dB.

En effet, outre l'éloignement du microphone (introduisant une dégradation "géométrique" du SNR de l'ordre de 6 dB par rapport à un kit filaire), le
10 comportement typique d'un porteur de casque, évalué sur un panel de personnes, montre que, d'une part, le niveau de voix, fort ou faible, qu'une personne adopte spontanément quand elle parle est modifié sensiblement lorsqu'elle porte un casque audio (par rapport à une situation où elle ne porte aucun casque) et que, d'autre part, ce niveau de voix dépend
15 aussi du type de casque porté (circum-aural, supra-aural ou intra-auriculaire), plus ou moins isolant.

Dans le cas d'une oreillette *Bluetooth*, le niveau de voix du porteur du casque pendant la conversation téléphonique est assez peu affecté par le port de cet appareil, car une seule oreille est obstruée alors que la se-
20 conde reste libre, ce qui permet au porteur du casque de percevoir à la fois le bruit environnant et sa propre voix de façon quasiment naturelle. En revanche, lorsque le porteur du casque porte un casque audio assez isolant du point de vue acoustique, il adopte généralement un niveau de voix notablement plus faible que la normale dans les situations bruyantes.
25 Ainsi, dans un contexte urbain de rue passante, le niveau de voix adopté peut être si faible qu'il en devient quasiment imperceptible pour une personne se trouvant à proximité (par exemple à un mètre), car le porteur du casque se perçoit comme étant dans une "bulle" et n'a plus le réflexe d'élever le niveau de sa voix pour la placer au-dessus du bruit ambiant de façon à être bien compris. Cette dégradation "comportementale" du SNR
30 peut être estimée à environ 4 à 6 dB de déficit acoustique dans les situations les plus typiques d'utilisation d'un casque nomade en contexte urbain.

Enfin, le niveau de voix du porteur du casque peut être affecté par le con-
35 texte "social" dans lequel il se trouve : en effet, dans une situation de

- promiscuité (métro, bureau,...) le niveau de voix du porteur du casque est modulé, par discrétion ou politesse, pour éviter de ne pas trop gêner les personnes à proximité. Cette diminution naturelle du niveau de voix vient s'ajouter aux déficits décrits plus haut, ce qui peut aboutir à un SNR
- 5 acoustique très dégradé, voisin de -10 dB, en entrée de l'algorithme de débruitage (on notera que cette dégradation "sociale" du SNR n'est pas spécifique à un casque audio, à la différence des deux autres dégradations mentionnées plus haut ("géométrique" et "comportementale"), mais elle vient aggraver un SNR déjà fortement dégradé).
- 10 De fait, à titre illustratif, le déficit cumulé de SNR est dans une situation typique compris entre 10 et 12 dB entre un système classique (kit filaire) et un casque fermé de type sans fil.
- Le problème de l'invention est lié notamment à l'utilisation d'un casque d'écoute sans fil qui utilise une technologie *Bluetooth* pour son couplage
- 15 avec un téléphone portable, car cette caractéristique rend impossible l'intégration du microphone au sein d'un câble comme cela se fait traditionnellement sur les casques filaires – donc sans bénéficier de ce placement avantageux pour la captation de la voix, et sans possibilité offerte au porteur du casque de rapprocher le microphone de sa bouche dans
- 20 une situation bruyante.
- Cette configuration fait dès lors apparaître une situation acoustique initiale dégradée, avec une dégradation "géométrique" du SNR acoustique de l'ordre de 6 dB environ, qui devra être rattrapée par l'algorithme de débruitage pour obtenir un niveau qualitatif de sortie similaire, voire supérieur, à celui d'un kit filaire.
- 25 D'autre part, le casque peut être muni d'un système dit de "contrôle actif de bruit" ou ANC (*Active Noise Control*), agissant sur le signal reproduit par les transducteurs du casque (que celui-ci soit ou non pourvu d'un microphone). Le principe de l'ANC consiste à capter la composante de bruit
- 30 ambiant et à superposer, temporellement et spatialement, à cette composante de bruit une onde acoustique qui est idéalement la copie inversée de l'onde de pression de la composante de bruit. Il s'agit de créer de cette manière une interférence destructive avec la composante de bruit et réduire, idéalement neutraliser, les variations de pression de l'onde acous-

tique parasite. Le EP 2 597 889 A1 (Parrot) décrit un casque muni d'un tel système ANC.

L'activation de l'ANC modifie le comportement du porteur du casque, en ce qu'il renforce l'effet "bulle" par une atténuation supplémentaire du bruit environnant perçu, avec une légère dégradation du SNR d'entrée par rapport au port d'un casque passif (sans ANC). Cette dégradation additionnelle reste toutefois modérée, de l'ordre de 1 dB environ, car l'activation de l'ANC a aussi pour effet secondaire d'atténuer le niveau perçu par le porteur du casque sur sa propre voix, en particulier dans les basses fréquences qui seront beaucoup moins sourdes, et de l'inciter à parler plus fort.

L'activation d'une autre technique, dite "anti-occlusion", ci-après "AOC", est également de nature à dégrader le SNR d'entrée par rapport au port d'un casque passif.

L'AOC (*Active Occlusion Cancellation*) est un mode de contrôle actif, différent de l'ANC, qui vise à effacer à la fois le phénomène d'occlusion entendu sur la voix et l'effet d'isolation du casque.

L'occlusion est le phénomène d'augmentation des basses fréquences, autour de 200 Hz, dans la perception de sa propre voix par l'utilisateur lorsque celui-ci porte un casque, la voix apparaissant beaucoup plus sourde qu'au naturel, ce qui crée une réelle gêne. En effet, en l'absence de casque, lorsque l'oreille n'est pas obstruée les sons de voix conduits par conduction osseuse au canal auditif sont très faiblement perçus, car ils sont évacués vers l'extérieur de l'oreille, qui présente l'impédance acoustique la plus faible par rapport à celle de la membrane tympanique. En revanche, lorsqu'un casque est porté, ce casque obstrue totalement ou partiellement le canal auditif, c'est-à-dire qu'il introduit une impédance acoustique importante à l'extrémité externe du canal auditif : cette impédance entraîne la mise en résonance au sein du canal auditif des sons transmis par conduction osseuse, et ainsi une amplification de la partie basse fréquence du signal de voix par rapport à une situation où le conduit auditif est ouvert, avec une élévation du niveau de l'ordre de 20 dB au-dessous de 500 Hz. Le porteur du casque perçoit alors sa voix de manière beaucoup plus sourde.

- Le but du traitement AOC est précisément de reproduire une perception des sons au plus proche du naturel, comme lorsqu'aucun casque n'est porté. En particulier, le porteur du casque souhaite parfois percevoir de manière naturelle sa propre voix, par exemple il souhaite pouvoir conver-
- 5 ser avec le locuteur distant dans une conversation téléphonique "mains-libres", ou avec une personne physiquement présente près de lui, en percevant sa propre voix de la même manière que s'il ne portait pas de casque. Dans d'autres situations, le porteur du casque souhaite percevoir parfaitement son environnement afin d'entendre par exemple la circulation
- 10 automobile, évaluer la distance des véhicules ou entendre un signal d'alarme, un message diffusé par le conducteur d'un transport en commun, etc.
- En d'autres termes, l'activation de l'AOC a pour but de transformer virtuellement, par des moyens purement électroniques et numériques, un
- 15 casque de type "fermé" pour simuler un casque "ouvert" avec suppression apparente du phénomène d'occlusion lorsque le porteur du casque parle et suppression active, à volonté, de l'isolation passive du casque.
- En ce qui concerne le SNR, l'activation de la fonction AOC anti-occlusion induit une légère augmentation du niveau de voix du locuteur, de l'ordre
- 20 de +1 à +2 dB. L'AOC efface en effet en partie l'effet "bulle" de l'ANC et remet le porteur du casque en contact avec son environnement, ce qui se traduit typiquement par une augmentation du SNR d'entrée, c'est d'ailleurs une raison pour laquelle le mode AOC, lorsqu'il est présent, est activé par défaut lors d'une communication téléphonique.
- 25 Toutefois, cette augmentation du niveau de voix du locuteur ne compense que très partiellement la dégradation "comportementale" du SNR décrite plus haut, et certains porteurs de casque préfèrent "rester dans leur bulle", c'est-à-dire en mode ANC sans AOC, et ne pas activer le mode AOC lors d'une communication téléphonique.
- 30 L'un dans l'autre, compte tenu des divers phénomènes exposés ci-dessus, on constate généralement, comme indiqué plus haut, un déficit cumulé du SNR compris entre -10 et -12 dB entre un système conventionnel dans la configuration la plus favorable (kit filaire) et un casque fermé de type sans fil tel que décrit plus haut, susceptible de mettre en
- 35 œuvre des fonctions de type ANC/AOC.

- Un but général de l'invention est de proposer une technique permettant, avec un tel casque, d'améliorer le SNR du signal capté par l'ensemble microphonique du casque, tout en préservant la qualité subjective du signal de voix du porteur de ce casque, de façon à délivrer en sortie un signal fidèle et intelligible transmis pour la communication téléphonique à un interlocuteur distant. Il s'agit également d'obtenir cette amélioration du SNR quelle que soit la situation de bruit dans laquelle se trouve le porteur du casque (bruit de rue, métro, train, bureau...) prise en compte par l'algorithme de traitement ANC.
- 5 Plus particulièrement, les algorithmes de débruitage du signal recueilli par le microphone du casque et destiné à être transmis au locuteur distant mettent en œuvre un "détecteur d'activité vocale" ou VAD (*Voice Activity Detector*) destiné à discriminer les périodes où le porteur du casque parle (donc avec un débruitage adapté pour ne pas déformer la voix et lui conserver toute son intelligibilité), et celles où il ne parle pas (où l'on peut alors appliquer un débruitage beaucoup plus agressif).
- 10 De fait, dans nombre de situations cette détection d'activité vocale est difficile à opérer ou conduit à des résultats imprécis.
- C'est d'ailleurs pour ces raisons que le résultat de l'analyse VAD n'est généralement pas donné sous forme binaire (parole présente ou parole absente), mais en termes de probabilité de présence de parole ou SPP (*Speech Presence Probability*), variant de façon continue sur un intervalle [0,1].
- 20 Plus précisément, la détermination de cette probabilité de présence de parole est effectuée sur d'étroites plages de fréquence, ou "bins fréquentiels", du signal recueilli, par exemple sur 256 bins fréquentiels d'une trame temporelle de signal après numérisation de celui-ci et analyse dans le domaine fréquentiel après transformation de Fourier. On dispose ainsi, pour une trame de signal donnée, de 256 valeurs de probabilité de présence de parole différentes, correspondant chacune à l'analyse de présence/absence de parole effectuée pour un bin fréquentiel spécifique.
- 30 De nombreuses techniques de détermination de probabilité de présence de parole ont été proposées, mais aucune ne permet d'obtenir des résultats de qualité constante quelles que soient les situations de bruit environnant susceptibles d'être rencontrées dans la réalité (la qualité des ré-
- 35

sultats peut être notamment quantifiée en termes de taux de vrais positifs par rapport au taux de faux positifs).

Ainsi, l'article de WH Shin et al. "Speech/non-Speech Classification Using Multiple Features for Robust Endpoint Detection", *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP '00 Proceedings*, Istanbul, 5-9 juin 2000, Vol. 3, pp. 1399-1402 décrit un détecteur d'activité vocale mettant en œuvre un classifieur à arbre de décision (algorithme CART). La technique décrite par ce document est toutefois une technique simplifiée dans laquelle :

- 10 – la détection d'activité vocale délivre simplement un indicateur binaire de type "tout ou rien" (présence vs. absence de parole) ;
- l'arbre de décision est purement binaire, l'aiguillage vers une branche ou une autre de l'arbre se faisant selon que la valeur binaire délivrée par chaque pré-classifieur est 0 ou 1 ; et
- 15 – les différents traitements en amont de l'arbre de décision sont opérés par des méthodes différentes, mais sur un même signal d'entrée.

Le but de l'invention est de proposer une nouvelle technique de détection d'activité vocale permettant de pallier les différents inconvénients exposés ci-dessus, afin de pouvoir délivrer des résultats avec en toutes circonstances une fiabilité très supérieure à ce qui a pu être obtenu jusqu'à présent.

Un autre but est de proposer un circuit amélioré de débruitage du signal, adapté pour utiliser spécifiquement les résultats du détecteur d'activité vocale de l'invention.

25 À cet effet, l'invention propose, plus précisément, un combiné micro-casque comprenant, de manière en elle-même connue :

- deux écouteurs d'oreille logeant chacun un transducteur de reproduction d'un signal audio ; et
- un réseau de détecteurs aptes à délivrer des signaux captés respectifs, chaque signal capté comportant une composante de bruit et le cas échéant une composante de parole du porteur du casque, le réseau de détecteurs comprenant :
 - 30 · une pluralité de détecteurs acoustiques de type microphonique placés à l'extérieur de la cavité acoustique ; et

- facultativement, au moins un détecteur non acoustique de type physiologique, apte à capter des vibrations vocales non acoustiques transmises par conduction osseuse interne du porteur du casque.
- 5 De façon caractéristique de l'invention, ce combiné micro-casque comprend en outre :
- des moyens d'analyse des signaux captés :
 - recevant sur des entrées respectives les signaux captés et/ou des combinaisons desdits signaux délivrés séparément par chacun
 - 10 desdits détecteurs du réseau de détecteurs,
 - opérant sur ces signaux et/ou combinaisons une pluralité de n traitements d'analyse distincts de détection d'activité vocale, et
 - délivrant en sortie une pluralité correspondante de n indicateurs élémentaires quantifiant une probabilité à valeur continue de présence de parole,
 - 15 l'ensemble des n indicateurs élémentaires formant un vecteur d'observation ; et
 - des moyens classifieurs du vecteur d'observation, aptes à :
 - comparer les n indicateurs élémentaires à valeur continue à un en-
 - 20 semble de seuils successifs prédéfinis hiérarchisés selon une configuration booléenne d'arbre de décision, et à
 - générer en réponse en sortie de l'arbre de décision un indicateur global de présence/absence de parole.
- En résumé, les moyens d'analyse de l'invention opèrent de telle manière
- 25 que (à la différence notamment de l'algorithme de Shin et al. mentionné plus haut) :
- les entrées de l'arbre soient des valeurs de probabilités à valeur continue sur l'intervalle $[0, 1]$, résultant d'une analyse d'un signal ;
 - l'arbre de décision soit un arbre formé de seuils hiérarchisés, l'arbre
 - 30 étant parcouru en comparant à chaque embranchement la valeur de probabilité de présence, résultat de l'analyse opérée en amont, avec un seuil correspondant ; et
 - les traitements préalables en amont de l'arbre de décision soient des traitements opérés sur des signaux d'entrée différents, en l'espèce des
 - 35 signaux délivrés par des micros différents du réseau de micros, ou des combinaisons différentes de ces micros du réseau, et non sur un même signal d'entrée par des méthodes différentes.

De plus, selon diverses caractéristiques subsidiaires avantageuses de l'invention :

- l'indicateur global de présence/absence de parole est un indicateur global quantifié, quantifiant une probabilité de présence de parole comprise entre deux extrema correspondant respectivement à une situation d'activité vocale avérée et à une situation d'absence d'activité vocale avérée. Il peut notamment être prévu à cet effet un comparateur à seuil, recevant en entrée l'indicateur global quantifié de présence/absence de parole et délivrant en sortie un indicateur global binaire de présence/absence de parole ;
- les signaux captés sont des signaux numérisés en trames successives avec, pour chaque trame, une décomposition en bins fréquentiels, et dans lequel les moyens d'analyse des signaux captés, les détecteurs d'activité vocale et les moyens classifieurs sont des moyens mis en œuvre pour chaque bin de chaque trame des signaux ;
- les moyens d'analyse comprennent des moyens d'analyse dans le domaine fréquentiel des signaux captés par tout ou partie des détecteurs acoustiques ou non-acoustiques du réseau de détecteurs, du groupe formé par : des moyens d'analyse de cohérence spatiale entre signaux captés par certains détecteurs acoustiques prédéterminés du réseau de détecteurs ; des moyens de quantification du déphasage entre signaux captés par certains détecteurs acoustiques prédéterminés du réseau de détecteurs ; des moyens d'estimation du rapport signal sur bruit des signaux captés et/ou combinaisons de signaux captés par certains détecteurs acoustiques prédéterminés du réseau de détecteurs ; des moyens probabilistes de détection d'activité vocale à partir des signaux captés par les détecteurs acoustiques ; et des moyens d'estimation du rapport signal sur bruit des signaux captés par le au moins un détecteur non acoustique du réseau de détecteurs ;
- les moyens d'analyse comprennent des moyens d'analyse dans le domaine temporel des trames successives des signaux captés par le au moins un détecteur non acoustique du réseau de détecteurs ;
- le casque comprend en outre des moyens de réduction de bruit, recevant en entrée les signaux captés et/ou des combinaisons desdits signaux captés et opérant sur ces signaux et/ou combinaisons de si-

gnaux un traitement paramétrable fonction de la valeur de l'indicateur global de présence/absence de parole délivré par les moyens classifieurs. Le traitement paramétrable peut notamment comprendre un traitement d'estimation d'une densité spectrale de bruit, sélectivement
5 mis en œuvre, ou non, en fonction d'une valeur booléenne de l'indicateur global de présence/absence de parole.

◇

10 On va maintenant décrire un exemple de mise en œuvre de la présente invention, en référence aux dessins annexés où les mêmes références désignent d'une figure à l'autre des éléments identiques ou fonctionnellement semblables.

La Figure 1 est une vue de face d'un premier type de micro-casque auquel est applicable l'invention, à savoir un micro-casque muni d'écouteurs
15 de type circumaural.

La Figure 2 est une vue en perspective du micro-casque de la Figure 1.

La Figure 3 est une vue de côté du casque de la Figure 1 en position sur la tête d'un porteur de ce casque, montrant plus précisément l'orientation
20 des axes de directivité du réseau de microphones par rapport à la bouche du porteur du casque.

La Figure 4 est une vue de face d'un second type de micro-casque auquel est applicable l'invention, à savoir un micro-casque de type sportif, muni d'écouteurs de type intra-auriculaire.

La Figure 5 est une vue de côté du casque de la Figure 4 en position sur la tête d'un porteur de ce casque, montrant plus précisément l'orientation
25 des axes de directivité du réseau de microphones par rapport à la bouche du porteur du casque.

La Figure 7 est une vue générale, sous forme de schéma par blocs, montrant et explicitant la manière dont les signaux issus des divers détecteurs acoustiques (microphoniques) et non acoustiques (accélérométriques) du micro-casque sont analysés, pour produire un certain nombre d'indicateurs élémentaires combinés par un classifieur de type arbre de décision, afin d'obtenir en sortie un indicateur global, quantifié ou binaire, de présence/absence de parole.
35

La Figure 8 est un diagramme de type ROC (taux de vrais positifs en fonction du taux de faux positifs) illustrant l'efficacité procurée par le classifieur de l'invention.

La Figure 9 illustre, sous forme de schéma par blocs, un réducteur de bruit utilisant les indicateurs globaux produits par l'arbre de décision de la Figure 7.

◇

On va maintenant décrire un exemple de mise en œuvre de la technique de l'invention.

De façon générale, l'invention est mise en œuvre par des moyens principalement logiciels, au moyen d'algorithmes appropriés exécutés par exemple par des microcontrôleurs ou processeurs numériques de signal. Pour la clarté de l'exposé, les divers traitements appliqués seront décomposés et schématisés par un certain nombre de blocs fonctionnels distincts présentés sous forme de circuits interconnectés, mais cette représentation n'a toutefois qu'un caractère illustratif, ces circuits comprenant des éléments communs et correspondant en pratique à une pluralité de fonctions globalement exécutées au sein d'un même logiciel.

Sur les Figures 1 à 3 on a représenté un combiné micro-casque comportant, de manière en elle-même classique, deux écouteurs 10, 10' réunis par un arceau de maintien 12. Chaque écouteur comprend une coque extérieure 14, 14' venant s'appliquer sur le contour de l'oreille du porteur du casque avec interposition d'un coussinet souple circumaural 16, 16' destiné à assurer une étanchéité satisfaisante, du point de vue acoustique, entre la région de l'oreille et l'environnement sonore extérieur.

Le micro-casque est muni d'un réseau formé d'une pluralité de microphones (ci-après également dénommés "détecteurs acoustiques"), dans l'exemple illustré au nombre de quatre, à savoir :

- un microphone 18 disposé en partie haute de l'écouteur gauche 10, délivrant un signal ci-après référencé *TopL* ;
- un microphone 18' situé en partie haute de l'écouteur droit 10', délivrant un signal désigné par la suite *TopR* ;
- un microphone 20 disposé en partie basse de l'écouteur gauche 10, délivrant un signal ci-après désigné *BotL* ; et

- un microphone 20' situé en partie basse de l'écouteur droit 10', délivrant un signal ci-après désigné *BotR*.

Les microphones 18 et 20 disposés sur l'écouteur gauche 10 (et de même les microphones 18' et 20' disposés sur l'écouteur droit 10') sont disposés l'un par rapport à l'autre, à une distance de l'ordre de 2 à 3 cm, de manière que l'axe Δ passant par les deux points où se trouvent ces microphones 18 et 20 soit dirigé vers la bouche du porteur du casque, en formant avec la direction D idéale passant par cette bouche un angle aussi réduit que possible.

Les microphones 18, 18', 20 et 20' sont avantageusement des microphones omnidirectionnels, qui seront donc amenés à capter aussi bien les ondes acoustiques produites par la bouche du porteur du casque lorsque celui-ci parle, que le bruit ambiant parasite qu'il s'agira d'éliminer dans les meilleures conditions possibles pour rendre intelligible le signal utile de voix du porteur du casque.

Le combiné micro-casque comprend également au moins un détecteur physiologique 22 de type accélérométrique ou autre (ci-après également dénommés "détecteur non acoustique"), par exemple disposé au niveau du coussinet 16' de manière à être appliqué contre la joue ou la tempe du porteur du casque avec un couplage le plus étroit possible. Il s'agit de recueillir au moyen de ce détecteur un signal transmis directement par voie osseuse (et non par voie acoustique aérienne comme dans le cas des détecteurs microphoniques), qui présente la caractéristique d'être, par nature, très peu corrompu par le bruit environnant. De manière en elle-même connue, le signal ainsi recueilli par ce détecteur non acoustique 22 permet, après filtrage et combinaison avec les signaux captés par les détecteurs acoustiques constitués par les microphones 18, 18', 20 et 20', de délivrer au système de communication un signal de parole du locuteur proche (le porteur du casque) dont l'intelligibilité aura été grandement améliorée.

Cette technique, ainsi que d'autres avantages procurés par le capteur physiologique (en particulier la possibilité d'utiliser le signal qu'il délivre pour calculer une fréquence de coupure d'un filtre dynamique), sont exposés notamment dans le EP 2 518 724 A1 (Parrot), auquel on pourra se référer pour de plus amples détails.

Les Figures 4 à 6 sont homologues des Figures 1 à 3 décrites ci-dessus, et montrent à titre illustratif un second type de micro-casque auquel est applicable l'invention, à savoir un micro-casque de type sportif, muni d'écouteurs de type intra-auriculaire. Les mêmes références y désignent des éléments identiques ou fonctionnellement semblables au micro-casque muni d'écouteurs de type circumaural des Figures 1 à 3.

La configuration d'ensemble d'un tel micro-casque de type sportif est par exemple celle du casque *Zik Sport* qui a été récemment présenté par la société Parrot, Paris, France.

En référence à la Figure 7, on va maintenant décrire la manière dont les signaux issus des divers détecteurs acoustiques (microphoniques) et non acoustiques (accélérométriques) du micro-casque sont analysés selon la technique spécifique de l'invention, pour produire un certain nombre d'indicateurs élémentaires combinés par un classifieur de type arbre de décision, afin d'obtenir en sortie un indicateur global, quantifié ou binaire, de présence/absence de parole.

Les signaux *TopL*, *BotL*, *TopR* et *BotR* produits par les $N = 4$ détecteurs acoustiques (les microphones 18, 18', 20 et 20') sont soumis à un filtrage passe-haut (HPF, blocs 100) de manière à en éliminer les composantes parasites de plus basse fréquence.

Les signaux *TopL*, *BotL*, *TopR* et *BotR* peuvent être combinés en paires, de manière à réaliser un certain nombre d'analyses spécifiques tenant compte de la position géométrique particulière sur le casque des microphones de la paire considérée.

Ainsi, un bloc 102 combine les signaux issus des microphones *TopL* et *BotL* pour une analyse de cohérence axiale côté gauche (L). Il s'agit d'une paire "axiale" en ce sens qu'elle regroupe deux microphones situés sur un même écouteur du casque (l'écouteur gauche en l'espèce) et dont l'axe commun Δ est orienté vers la bouche du porteur du casque (cf. Figures 3 et 6).

Les signaux combinés de la paire axiale $\{TopL, BotL\}$ sont soumis par exemple à un traitement de quantification du déphasage du signal capté, évalué entre le microphone *TopL* et le microphone *BotL* (bloc 104), et à un autre traitement d'estimation du rapport signal/bruit (bloc 106) sur cette paire axiale de microphones situés du côté gauche.

En ce qui concerne le traitement opéré par le bloc 106, on pourra notamment se référer aux algorithmes décrits dans le WO 2007/099222 A1 (Parrot SA), qui sont basés sur une estimation du bruit via l'utilisation d'un algorithme MCRA (*Minima Controlled Recursive Averaging*), moyennage
5 récursif par contrôle des minima, par exemple décrit par I. Cohen et B. Berdugo, Noise Estimation by Minima Controlled Recursive Averaging for Robust Speech Enhancement, *IEEE Signal Processing Letters*, Vol. 9, No 1, pp. 12-15, Jan. 2002), avec une hypothèse de bruit stationnaire pour déduire une probabilité de présence/absence de parole. Le SNR est calculé *a posteriori* comme le ratio de la norme du signal bruité capté par les
10 microphones et de l'estimation du bruit délivré par l'algorithme MCRA. On pourra se référer au document précité pour de plus amples détails sur la manière d'opérer ce traitement.

Outre les paires axiales, il est également possible de considérer des paires "transverses" de microphones, correspondant à des microphones
15 symétriques dont l'un se situe côté gauche et l'autre côté droit, par exemple la paire $\{TopL, TopR\}$ et/ou la paire $\{BotL, BotR\}$.

Les signaux de ces paires, combinés par le bloc 108, font l'objet par exemple d'un traitement d'analyse de quantification du déphasage du signal capté, évalué entre le côté gauche et le côté droit (bloc 110), et d'un
20 autre traitement d'estimation du rapport signal/bruit (bloc 112) à partir des signaux recueillis par ces paires transverses.

Il est également prévu de combiner (bloc 114) les signaux de la paire axiale $\{TopR, BotR\}$, par un traitement de quantification du déphasage du signal capté, évalué entre le microphone *TopR* et le microphone *BotR*
25 (bloc 116), et un autre traitement d'estimation du rapport signal/bruit (bloc 118) sur cette paire axiale de microphones situés du côté droit.

Les traitements de signal opérés côté droit par les blocs 116 et 118 sont les mêmes que ceux, décrits plus hauts, opérés côté gauche par les blocs
30 104 et 106.

Enfin, les signaux *TopL*, *BotL*, *TopR* et *BotR* issus des quatre microphones 18, 18', 20 et 20' sont également appliqués ensemble à un circuit 120 du type détecteur d'activité vocale probabiliste, par exemple mettant en œuvre un algorithme de type MCRA.

Une telle détection, en elle-même connue, est basée sur l'estimation de l'énergie de la composante de bruit pseudo-stationnaire dans les signaux captés par les microphones, qui est évaluée par un algorithme du type MCRA.

- 5 En ce qui concerne le capteur physiologique (accéléromètre 22), après traitement par un filtre passe-haut 122, le signal délivré est soumis à divers traitements d'analyse (bloc 124) comprenant par exemple :
- un traitement d'estimation du rapport signal/bruit du signal accélérométrique dans les basses fréquences correspondant à la gamme de
 - 10 fonctionnement utile de l'accéléromètre, à savoir les fréquences inférieures à 1500 Hz (bloc 126) ;
 - un traitement d'estimation du rapport signal/bruit du signal accélérométrique en bande étendue (bloc 128) ; et
 - une analyse temporelle des trames successives (bloc 130).
- 15 En ce qui concerne le bloc 126, on notera que le traitement qu'il opère de mesure du SNR du signal accélérométrique est principalement une mesure de SNR vis-à-vis d'un bruit électrique, du fait que le bruit présent sur le signal accélérométrique est essentiellement un bruit de souffle électrique lié au capteur (bruit thermique) et à la numérisation du signal (bruit
- 20 de quantification). Ce bruit électrique étant stationnaire, il est aisé d'obtenir une évaluation du SNR.
- Toutefois, lorsque le bruit acoustique environnant est très élevé (par exemple un bruit de marteau-piqueur dans l'environnement immédiat du porteur du casque) et qu'il entraîne une légère vibration physique de la
- 25 joue du locuteur, on peut observer une composante de bruit acoustique parasite qui dépasse le bruit de souffle électrique. Pour tenir compte de cette éventuelle situation, l'algorithme d'évaluation du SNR prévoit de mettre à jour l'estimation du plancher de bruit pour tenir compte de l'élévation de ce même plancher de bruit en cas de bruit acoustique fort. On
- 30 obtient ainsi en sortie du bloc 126 une détection d'activité dans la plage de fonctionnement de l'accéléromètre, à savoir pour des fréquences inférieures à 1500 Hz.
- En ce qui concerne le bloc 128, le traitement qu'il opère tient compte du fait qu'il est possible d'étendre artificiellement la gamme de fonctionne-
- 35 ment du capteur accélérométrique au-delà de la limite de 1500 Hz, dans

une plage correspondant aux harmoniques de rang faible de la voix, par une technique d'extension de bande, notamment par l'application d'une fonction non linéaire de génération d'harmoniques.

Dans le cas présent, cette technique d'extension de bande est applicable car le niveau de bruit est faible sur le capteur, et l'on peut atteindre une extension typiquement jusqu'à une fréquence de 3 kHz. L'estimation du SNR est ensuite opérée de la même manière que précédemment dans le domaine des basses fréquences. On dispose ainsi d'un détecteur d'activité vocale supplémentaire, dont le domaine d'analyse est étendu dans le médium.

En ce qui concerne le bloc 130, on notera que le traitement qu'il effectue est une analyse opérée dans le domaine temporel, à la différence des analyses opérées par les blocs 104, 106, 110, 112, 116, 118, 120, 126 et 128, qui sont toutes opérées dans le domaine fréquentiel.

Cette analyse dans le domaine temporel répartit les trames de signal successives reçues en deux classes, selon que l'algorithme détecte :

- i) un signal de parole voisé présent dans le signal capté (les signaux voisés sont ceux qui impliquent une vibration des cordes vocales, qui est transmise vers la joue, à la différence des signaux vocaux non voisés correspondant à des émissions sonores fricatives ou plosives), ou
- ii) du bruit seul ou bien un mélange de bruit et de signal de parole non voisé.

Cette discrimination en deux classes est en particulier utile dans les basses fréquences, où l'on peut observer un SNR très faible. L'absence de parole voisée permet à l'algorithme de mettre à jour son estimation de bruit de façon réactive, et sans risque sur la parole. Cette analyse est basée sur l'hypothèse que si l'on détecte des signaux vocaux dans les fréquences basses, il est très probable qu'on en retrouve également dans les fréquences plus élevées.

En résumé :

- on dispose d'un réseau de N détecteurs acoustiques et M détecteurs non acoustiques délivrant simultanément des signaux respectifs tous différents. Dans l'exemple illustré on a $N = 4$ détecteurs acoustiques, à savoir les microphones 18, 18', 20 et 20' délivrant les signaux *TopL*,

- BotL*, *TopR* et *BotR*, et $M = 1$ détecteur non acoustique, à savoir le capteur physiologique accélérométrique 22 délivrant le signal *Acc* ; et
- sur ces $N+M$ signaux on opère parallèlement n traitements d'analyse différents. Dans l'exemple illustré, on a $n = 10$ traitements différents,
- 5 correspondant aux blocs 104, 106, 110, 112, 116, 118, 120, 126, 128 et 130.

On notera que ces n traitements d'analyse sont opérés distinctement et simultanément sur chaque bande de fréquence ou "bin" fréquentiel de chaque trame successive de signal (une trame comprenant typiquement

10 256 bins).

Ces n traitements différents produisent une batterie de n indicateurs élémentaires respectifs donnant chacun, pour le bin considéré, une valeur correspondante de probabilité de présence de parole (n détections d'activité vocale schématisées par les blocs 140).

15 On évalue ainsi, pour chaque bin de chaque trame, une pluralité d'indicateurs de présence/absence de parole différents, indicateurs :

- qui sont tous complémentaires, car ils sont établis chacun à partir de caractéristiques différentes et de traitements différents : analyse de cohérence axiale ou transversale de détecteurs acoustiques, estimations du SNR selon différentes techniques et à partir d'informations différentes, analyses d'un signal non acoustique selon plusieurs aspects différents, etc.,
- 20
- mais qui sont tous imparfaits, et ne peuvent donc être utilisés individuellement pour décider en toutes circonstances et avec un degré
- 25 élevé de vraisemblance de la présence ou non d'un signal de parole à un instant donné dans un bin fréquentiel donné.

Dans l'exemple illustré, on a $n = 10$ informations différentes pour chaque bin fréquentiel de chaque trame temporelle, ces informations étant issues des signaux provenant de quatre microphones et d'un accéléromètre. Ces

30 informations auront une fiabilité variable selon le contexte de bruit présent. En effet, certains détecteurs permettent de distinguer efficacement la voix d'un bruit directif par des propriétés de direction d'arrivée, d'autres détecteurs sont spécialisés pour détecter voix et bruit dans les basses fréquences à SNR faible, d'autres encore sont spécialisés pour séparer la

35 voix d'un champ de bruits diffus, etc. De fait, les détecteurs s'appuient sur

des hypothèses distinctes et leur pertinence est variable selon la situation de bruit et les hypothèses qui sont effectivement vérifiées.

Le but de l'invention est, à partir de ces n indicateurs élémentaires tous complémentaires et tous imparfaits, de produire un indicateur global
5 unique de fiabilité élevée, c'est-à-dire présentant un très faible taux de faux positifs.

L'idée de base de l'invention consiste à appliquer concurremment ces n indicateurs élémentaires de probabilité de présence de parole à un classifieur supervisé de type arbre de décision 142, pour obtenir *in fine* un
10 indicateur global unique de présence/absence de parole, pour chaque bin fréquentiel et chaque trame du signal recueilli.

L'arbre de décision 142 utilise comme information d'entrée le "vecteur d'observation" constitué des n indicateurs élémentaires de probabilité de présence de parole produits par les n traitements décrits plus haut. On
15 notera que ces informations d'entrée sont obtenues par des traitements en eux-mêmes connus (traitements des blocs 102 à 130).

La classification supervisée par un arbre de décision 142 permet de réduire la détection de l'activité vocale à une cascade de comparaisons à des seuils 144 prédéfinis respectifs, la classification étant opérée entre
20 deux classes (Classe 1 = voix absente ; Classe 2 = voix présente).

Ce procédé de classification supervisée peut également permettre de sélectionner parmi les n traitements d'analyse différents ceux qui sont les plus pertinents sur chaque bande de fréquence, évitant ainsi d'exécuter la totalité de ces traitements sur la totalité des bins fréquentiels. On peut
25 ainsi obtenir des sous-ensembles classifieurs différenciés et spécialisés selon le domaine fréquentiel.

L'arbre de décision est une méthode, en elle-même bien connue dans d'autres domaines, basés sur un algorithme dans lequel une multiplicité de données d'entrées (dans le cas présent, les n indicateurs élémentaires de probabilité de présence de parole délivrés par les n détecteurs d'activité
30 vocale 140 en sortie des n traitements d'analyse respectifs différents) sont comparées à des seuils respectifs hiérarchisés entre eux.

Un tel algorithme peut être formalisé par une équation booléenne comprenant une succession de termes (correspondant aux comparaisons des
35 différents paramètres par rapport aux seuils respectifs) qui sont combinés

par une suite de ET logiques, et dont le résultat est un résultat global unique.

Dans le cas présent, les différents seuils 144 de l'arbre de décision 142 sont des seuils déterminés préalablement à partir d'une base d'apprentissage comprenant notamment une base de bruits réels captés dans des conditions représentatives de l'usage possible du micro-casque. Des enregistrements sont réalisés avec l'ensemble des capteurs acoustiques et non acoustiques, puis mixés à des signaux de voix obtenus en l'absence de bruit (dans une salle anéchoïque) pour une diversité d'utilisateurs possibles, et avec différents niveaux de SNR.

Une fois l'apprentissage statistique ainsi réalisé, les valeurs des seuils obtenues sont introduites dans l'arbre de décision, où ces seuils sont conservés dans une mémoire et figés.

On notera que l'on utilise un arbre de décision indépendant, et de profondeur différente, pour chaque bin fréquentiel.

On dispose ainsi d'autant d'arbres de décision que de bins fréquentiels, par exemple 256 arbres indépendants dans le cas où chaque trame est divisée en 256 bins.

Chaque arbre de décision délivre en sortie 146 une valeur 148 d'indicateur global de probabilité de présence de parole $P_{supervised}(k,l)$ pour le bin l de la trame de rang k . La valeur $P_{supervised}(k,l)$ de cet indicateur global (ci-après dénommé "indicateur global quantifié" de présence de parole) est une probabilité à valeur continue sur l'intervalle $[0,1]$, la valeur 0 correspondant à une absence certaine de parole et la valeur 1 correspondant à une présence certaine de parole.

La valeur $P_{supervised}(k,l)$ (probabilité à valeur continue) donnée par l'arbre de décision est appliquée à un comparateur à seuil 150 pour délivrer en sortie 152 une valeur binaire $VAD_{supervised}(k,l)$ de probabilité de présence de parole, ci-après dénommée "indicateur global binaire" de présence de parole), ceci pour chaque bin de chaque trame successive. Le seuil P1 du comparateur est paramétrable (le paramètre étant prédéterminé et conservé dans une mémoire, ce paramètre étant éventuellement réajustable par le porteur du casque), typiquement $P1 = 0,5$.

La Figure 8 illustre les performances du détecteur d'activité vocale de l'invention, basé sur une probabilité supervisée par classifieur à arbre de décision.

Cette Figure 8 est une représentation de type ROC (*Receiver Operating Characteristic*) quantifiant la performance sous forme d'une courbe caractéristique taux de vrais positifs/taux de faux positifs, la performance étant d'autant meilleure que l'aire sous la courbe est élevée (à comparer à une décision purement aléatoire, correspondant à la courbe A).

On a reporté sur ce diagramme ROC en B1 et B2 les résultats respectifs obtenus par deux détecteurs individuels d'activité vocale (correspondant à deux types de traitements distincts) opérant indépendamment, et en C le résultat obtenu par fusion de ces deux détecteurs au moyen du classifieur à arbre de décision selon l'invention.

Cet exemple, donné pour deux détecteurs élémentaires, montre une amélioration significative (forte diminution du taux de faux positifs) par rapport à l'un ou l'autre des détecteurs pris isolément. On comprendra que, pour un plus grand nombre de détecteurs (jusqu'à dix détecteurs dans l'exemple illustré Figure 7), cette amélioration sera d'autant plus élevée que le nombre de détecteurs élémentaires différents est important, chacun étant susceptible de prendre en compte une situation particulière d'environnement de bruit rencontrée par le porteur du micro-casque.

Sur la Figure 9, on a illustré une application particulière du détecteur d'activité vocale de l'invention à un traitement amélioré de débruitage du signal, adapté pour utiliser spécifiquement les résultats du détecteur d'activité vocale de l'invention.

Le but du traitement illustré Figure 9 est d'améliorer le SNR d'un signal d'entrée $y(t)$ 154, par exemple un signal microphonique destiné à être utilisé comme signal de locuteur proche dans le cadre d'une conversation téléphonique, tout en apportant beaucoup de soin à la préservation d'une qualité d'écoute subjectivement satisfaisante.

Le signal d'entrée $y(t)$ est segmenté en trames temporelles successives (bloc 156), et chaque trame temporelle $y(k)$ est analysée dans le domaine fréquentiel par transformation rapide de Fourier (bloc 158), pour donner une série de signaux $y(k,l)$ correspondant à l bins fréquentiels, par exemple $l = 256$ bins, pour chaque trame de rang k . Le signal $y(k,l)$ obte-

nu fait l'objet (branche 160) d'un traitement d'évaluation de la norme du signal (bloc 162), dont la sortie 164 est appliquée à un estimateur de bruit (bloc 168).

De façon caractéristique, la probabilité supervisée estimée en sortie de l'arbre de décision (valeur $P_{supervised}(k,l)$ à valeur continue 148 et valeur $VAD_{supervised}(k,l)$ binaire 152 de probabilité de présence de parole) et appliquée au réducteur de bruit de la Figure 9 est utilisée en entrée d'un étage de réduction de type MMSE-OMLSA (*Minimum Mean-Squared Error - Optimally-Modified Log Spectral Amplitude*) en lui-même de type connu, et par exemple décrit dans les WO 2007/099222 A1 (Parrot SA) ou EP 2 772 916 A1 (Parrot SA), auxquels on pourra se référer pour de plus amples détails.

Plus précisément, dans la structure classique de débruitage OM-LSA, une grandeur essentielle est l'estimation de la densité spectrale de puissance du bruit $PSD_{Noise}(k,l)$, car cette densité est utilisée pour estimer le SNR *a priori* $SNR_{prio}(k,l)$ et le SNR *a posteriori* $SNR_{post}(k,l)$, grandeurs qui sont au cœur de l'algorithme OM-LSA (la densité spectrale de bruit intervenant au dénominateur de ces deux grandeurs).

Du fait de la fiabilité de la probabilité de parole supervisée obtenue par la mise en oeuvre de l'invention, il est possible d'élaborer un nouveau modèle d'estimation de $PSD_{Noise}(k,l)$. En effet, la précision accrue de la probabilité de présence de parole, issue des $N+M$ détecteurs d'activité vocale et de leur fusion supervisée, par rapport à une probabilité de présence plus classique (monocanal ou déduite de capteurs acoustiques uniquement) remet en cause le calcul de $PSD_{Noise}(k,l)$. Concrètement, on peut effectuer une mise à jour plus directe de $PSD_{Noise}(k,l)$ sur la base de la probabilité supervisée, et ainsi éliminer certains mécanismes classiques de protection et de lissage, afin d'obtenir le maximum de fidélité et de détail dans l'estimation du bruit.

On commence par définir, pour chaque bin fréquentiel et chaque trame, un facteur de lissage $\alpha_{psd}(k,l) = \alpha + (1-\alpha) * P_{supervised}(k,l)$, où $P_{supervised}(k,l)$ est la probabilité supervisée à valeur continue et α est un facteur de lissage fixe compris entre 0 et 1 (typiquement 0,85).

α_{psd} est donc grand si la probabilité est élevée, ce qui permet de ne pas tenir compte de la norme courante du signal utile. *A contrario*, si la proba-

bilité est faible, on met à jour l'estimation du bruit sur la norme courante, en appliquant la formule :

$$PSDNoise(k,l) = \alpha_{psd}(k,l) * PSDNoise(k,l-1) + (1 - \alpha_{psd}(k,l)) * NormInput(k,l),$$

5

NormInput(k,l) étant la norme lissée de la transformée de Fourier du signal acoustique d'entrée *Input(t)*.

De façon caractéristique de l'invention, la valeur binaire de probabilité de présence de parole $VAD_{supervised}(k,l)$ est également prise en compte. En effet, cette valeur est très fiable (taux de faux positifs compris entre 1 % et 5 %), ce qui permet de réaliser une mise à jour intégrale de l'estimation du bruit sur la norme courante du signal d'entrée lorsque $VAD_{supervised}(k,l) = 0$ (absence de parole). Ceci revient à forcer $\alpha_{psd}(k,l)$ à zéro dans la formule précédente. Ainsi, si $VAD_{supervised}(k,l) = 0$, alors $\alpha_{psd}(k,l) = 0$ et

15 $PSDNoise(k,l) = NormInput(k,l)$.

Cette modification simple a beaucoup d'impact, car concrètement les valeurs continues de $P_{supervised}(k,l)$ ne seront utilisées que lorsque le seuil $P1$ sera dépassé ; dans le cas inverse, on peut effectuer une mise à jour instantanée qui permet de récupérer tous les détails du bruit et de son "grain".

20

Il faut signaler qu'il peut toutefois être avantageux d'appliquer un lissage fréquentiel sur $P_{supervised}(k,l)$ avant le calcul de $VAD_{supervised}(k,l)$, de façon à renforcer encore la robustesse de ce détecteur (ceci n'ayant pas d'impact sur le caractère instantané de la mise à jour décrite ci-dessus).

25 La manière de procéder, particulière à l'invention, que l'on vient de décrire est schématisée sur la Figure 9 par le fait que, selon la valeur de l'indicateur global binaire $VAD_{supervised}(k,l)$, on met en œuvre l'estimation du bruit de façon conventionnelle par calcul de la densité spectrale (sortie 170 du bloc 168), ou non, comme cela est schématisé par le commutateur

30 172 commandé par la valeur binaire $VAD_{supervised}(k,l)$. Si $VAD_{supervised}(k,l) = 0$, alors le traitement du bloc 168 est remplacé par l'application du signal 164 directement issu du bloc 162 de calcul de la norme du signal utile. En effet, dans la mesure où l'algorithme retient l'hypothèse qu'il n'y a pas de parole dans le signal recueilli, il peut décider que le spectre du bruit cor-

respond au spectre du signal d'entrée 164, et qu'il n'est pas nécessaire de procéder à une estimation de ce bruit par le bloc 168.

Les étapes suivantes du débruitage sont réalisées de manière en elle-même conventionnelle, avec un étage de calcul de gain OM-LSA (bloc 5 174) recevant d'une part le signal $y(k,l)$ (branche 176) et d'autre part l'estimation du bruit $PSD_{Noise}(k,l)$ calculée de la façon que l'on a décrite plus haut, c'est-à-dire en tirant avantageusement parti de l'indicateur global binaire $VAD_{supervised}(k,l)$ en sus de l'indicateur quantifié $P_{supervised}(k,l)$.

De façon en elle-même connue, un algorithme OM-LSA améliore le calcul 10 du gain LSA (*Log-Spectral Amplitude*) en le pondérant par la probabilité de présence de parole, la réduction de bruit appliquée étant d'autant plus importante que la probabilité de présence de parole est faible, c'est-à-dire que le gain appliqué sera d'autant plus faible que la probabilité de présence de parole est faible.

15 On pourra se référer pour de plus amples détails à la description du EP 2 772 916 A1 (Parrot) qui décrit un procédé de débruitage mettant en œuvre un tel type de traitement de réduction du bruit avec un algorithme OM-LSA.

Une autre donnée d'entrée du traitement opéré par l'étage 174 de calcul 20 du gain OM-LSA est un paramètre G_{min} dénommé "dureté de débruitage", déterminé en fonction de critères subjectifs obtenus lors de tests d'écoute. Il peut être avantageux de disposer, comme illustré en 178, de deux réglages différents de dureté de débruitage, selon que la fréquence d'échantillonnage est de type *narrowband* à 8 kHz ou *wideband* à 16 kHz.

25 Dans le premier cas (*narrowband*) l'expérience montre qu'un débruitage agressif, typiquement de -21 dB, est indispensable pour faire ressortir la voix d'une quantité de bruit élevée et valoriser l'intelligibilité, au prix d'une distorsion sur le timbre sur ce signal qui aura de toute façon une fidélité de restitution limitée par la fréquence d'échantillonnage moindre. Le problème se pose de façon différente en mode *wideband*, car le signal d'en- 30 trée est naturellement de meilleure qualité et plus intelligible, même en présence de bruit, en raison de l'ajout de la partie des fréquences les plus élevées dans le domaine 4 kHz-8 kHz, qui comporte souvent moins de bruit et apporte beaucoup de détails dans la voix. Dans ce mode, un dé- 35 bruitage plus modéré, typiquement de -15 dB, est plus approprié, car les

défauts introduits par le débruitage sur la fidélité du timbre seront moins bien tolérés par le locuteur distant qu'en mode *narrowband*.

Le gain $G_{OMLSA}(k,l)$ calculé par l'étage 174 est appliqué à un étage 182 recevant le signal $y(k,l)$ (branche 184), pour donner un signal débruité 186.

- 5 Ce signal est soumis à une transformation rapide de Fourier inverse 188 pour reconstituer le signal 190 dans le domaine temporel, puis les trames initialement segmentées sont synthétisées (bloc 192) pour donner le signal de sortie 194 reflétant le signal initial d'entrée 154 après traitement de débruitage.
-

REVENDICATIONS

1. Un casque audio de type combiné micro-casque, comprenant :
 - deux écouteurs d'oreille (10, 10') logeant chacun un transducteur de reproduction d'un signal audio ; et
 - un réseau de détecteurs aptes à délivrer des signaux captés respectifs, chaque signal capté comportant une composante de bruit et le cas échéant une composante de parole du porteur du casque, le réseau de détecteurs comprenant :
 - une pluralité de détecteurs acoustiques (18, 18', 20, 20') de type microphonique placés à l'extérieur de la cavité acoustique ; et
 - facultativement, au moins un détecteur non acoustique (22) de type physiologique, apte à capter des vibrations vocales non acoustiques transmises par conduction osseuse interne du porteur du casque,
- caractérisé en ce qu'il comprend en outre :
 - des moyens d'analyse des signaux captés (*TopL*, *BotL*, *TopR*, *BotR*, *Acc*) :
 - recevant sur des entrées respectives les signaux captés et/ou des combinaisons desdits signaux délivrés séparément par chacun desdits détecteurs du réseau de détecteurs,
 - opérant sur ces signaux et/ou combinaisons une pluralité de n traitements d'analyse distincts (104, 106, 110, 112, 116, 118, 120, 126, 128, 130) de détection d'activité vocale, et
 - délivrant en sortie une pluralité correspondante de n indicateurs élémentaires quantifiant une probabilité à valeur continue de présence de parole,
 - l'ensemble des n indicateurs élémentaires formant un vecteur d'observation ; et
 - des moyens classifieurs (142, 144) du vecteur d'observation, aptes à :
 - comparer les n indicateurs élémentaires à valeur continue à un ensemble de seuils successifs prédéfinis (144) hiérarchisés selon une configuration booléenne d'arbre de décision (142), et à
 - générer en réponse en sortie de l'arbre de décision un indicateur global de présence/absence de parole (148).

2. Le casque de la revendication 1, dans lequel l'indicateur global de présence/absence de parole est un indicateur global quantifié ($P_{\text{supervised}}(k,l)$), quantifiant une probabilité de présence de parole comprise entre deux extrema correspondant respectivement à une situation d'activité vocale avérée et à une situation d'absence d'activité vocale avérée.
- 5
3. Le casque de la revendication 2, comprenant en outre :
- un comparateur (150) à seuil (P_1), recevant en entrée l'indicateur global quantifié de présence/absence de parole (148) et délivrant en sortie un indicateur global binaire ($VAD_{\text{supervised}}(k,l)$), de présence/absence de parole.
- 10
4. Le casque de la revendication 1, dans lequel les signaux captés sont des signaux numérisés en trames successives avec, pour chaque trame, une décomposition en bins fréquentiels, et dans lequel les moyens d'analyse des signaux captés, les détecteurs d'activité vocale et les moyens classifieurs sont des moyens mis en œuvre pour chaque bin de chaque trame des signaux.
- 15
5. Le casque de la revendication 1, dans lequel les moyens d'analyse comprennent des moyens d'analyse dans le domaine fréquentiel des signaux captés par tout ou partie des détecteurs acoustiques ou non-acoustiques du réseau de détecteurs, du groupe formé par : des moyens (102-118) d'analyse de cohérence spatiale entre signaux captés par certains détecteurs acoustiques prédéterminés du réseau de détecteurs ; des moyens (104, 110, 116) de quantification du déphasage entre signaux captés par certains détecteurs acoustiques prédéterminés du réseau de détecteurs ; des moyens (106, 118) d'estimation du rapport signal sur bruit des signaux captés et/ou combinaisons de signaux captés par certains détecteurs acoustiques prédéterminés du réseau de détecteurs ; des moyens (120) probabilistes de détection d'activité vocale à partir des signaux captés par les détecteurs acoustiques ; et des moyens (126, 128) d'estimation du rapport signal sur bruit des signaux captés par le au moins un détecteur non acoustique du réseau de détecteurs.
- 20
- 25
- 30
- 35
6. Le casque de la revendication 1, dans lequel les moyens d'analyse comprennent des moyens (130) d'analyse dans le domaine temporel des

trames successives des signaux captés par le au moins un détecteur non acoustique du réseau de détecteurs.

7. Le casque de la revendication 1, comprenant en outre :
- 5 – des moyens de réduction de bruit (156-192), recevant en entrée les signaux captés et/ou des combinaisons desdits signaux captés et opérant sur ces signaux et/ou combinaisons de signaux un traitement paramétrable fonction de la valeur de l'indicateur global de présence/absence de parole ($P_{\text{supervised}}(k,l)$ (148) ; $VAD_{\text{supervised}}(k,l)$ (152))
- 10 délivré par les moyens classifieurs.
8. Le casque de la revendication 7, dans lequel ledit traitement paramétrable comprend un traitement d'estimation d'une densité spectrale de bruit (168), sélectivement mis en œuvre, ou non, en fonction d'une
- 15 valeur booléenne de l'indicateur global de présence/absence de parole ($VAD_{\text{supervised}}(k,l)$).
-

1/4

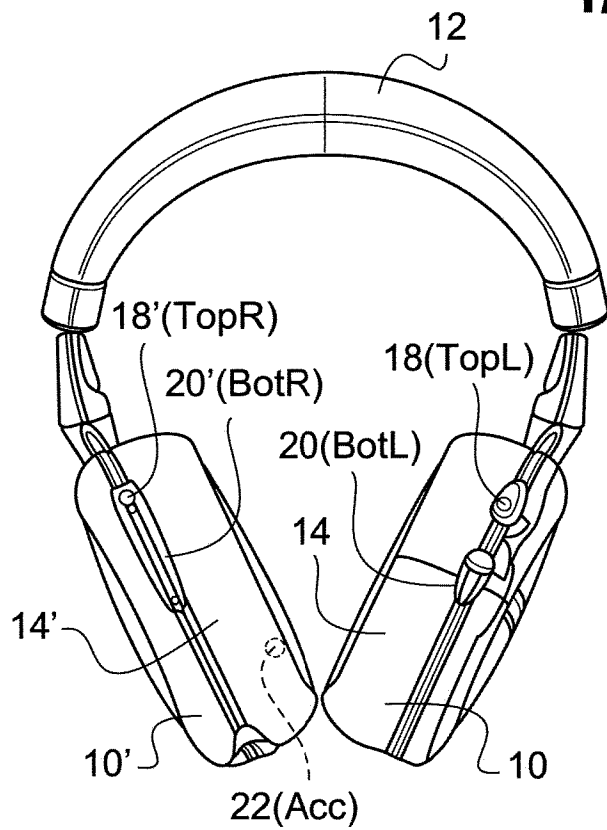


Fig.1

Fig.2

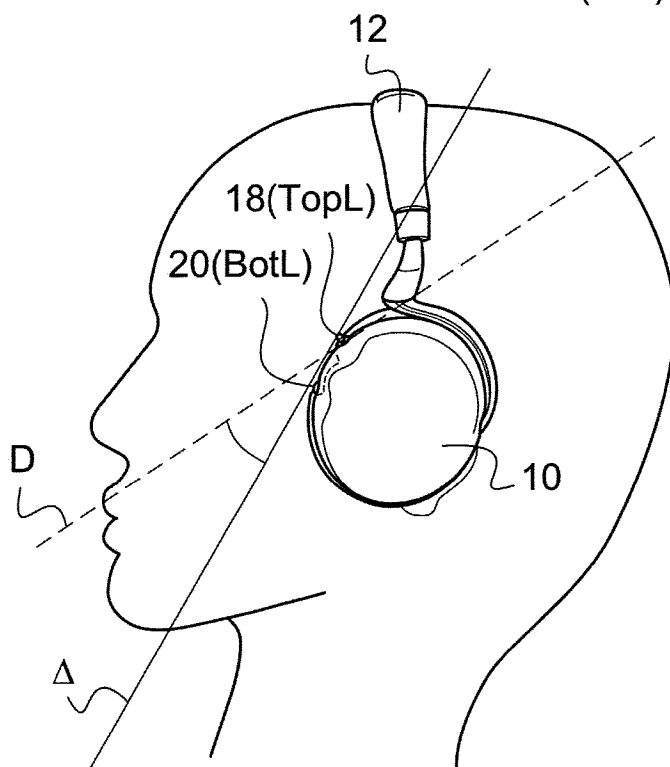
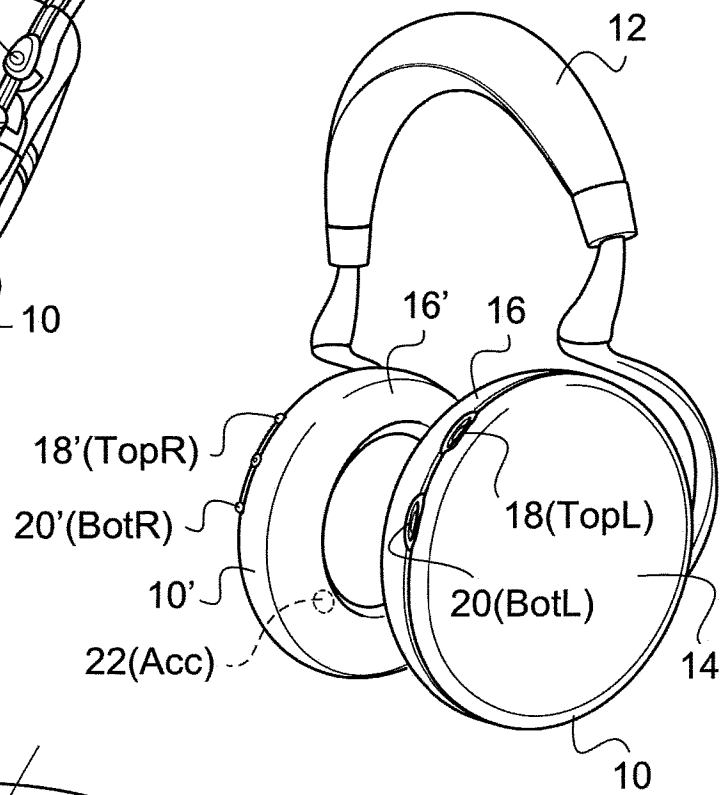
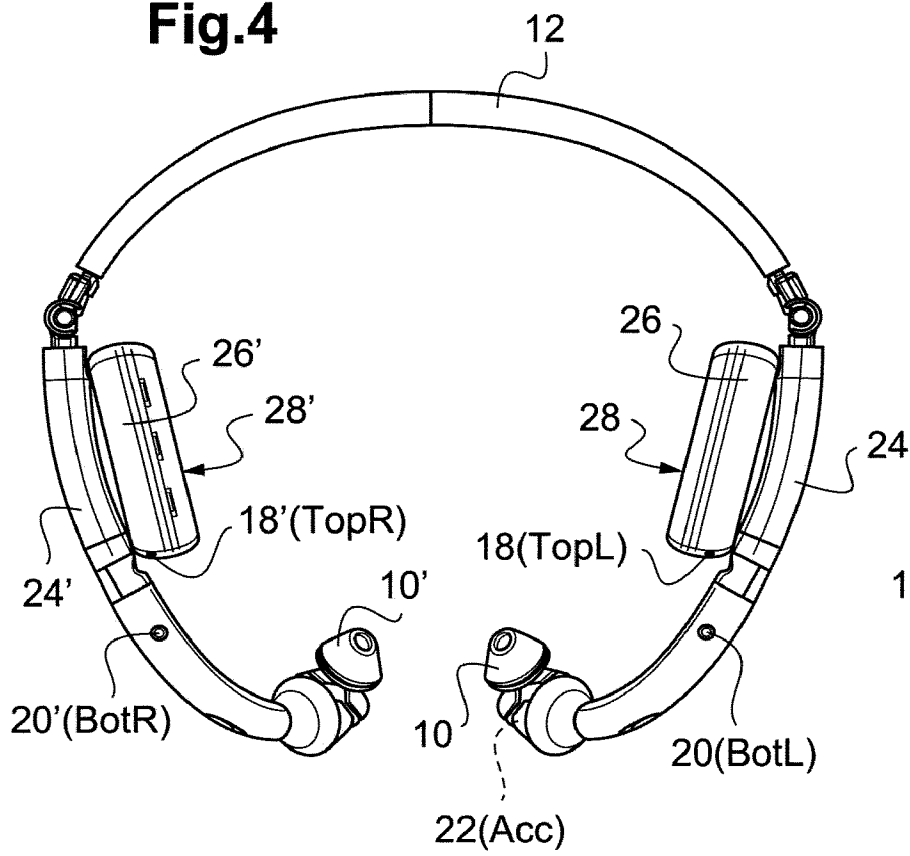
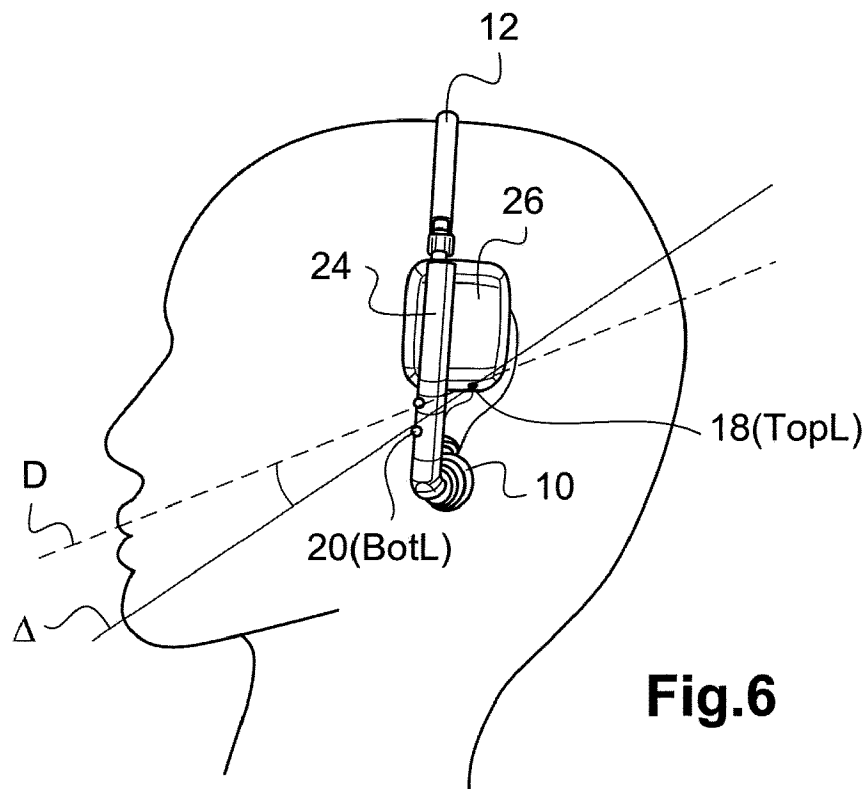
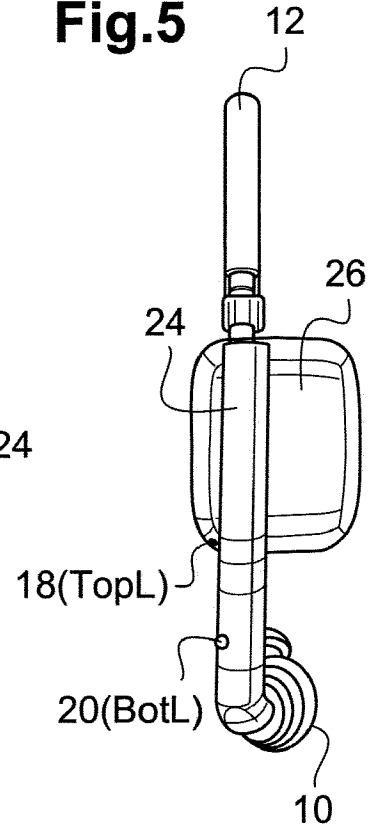
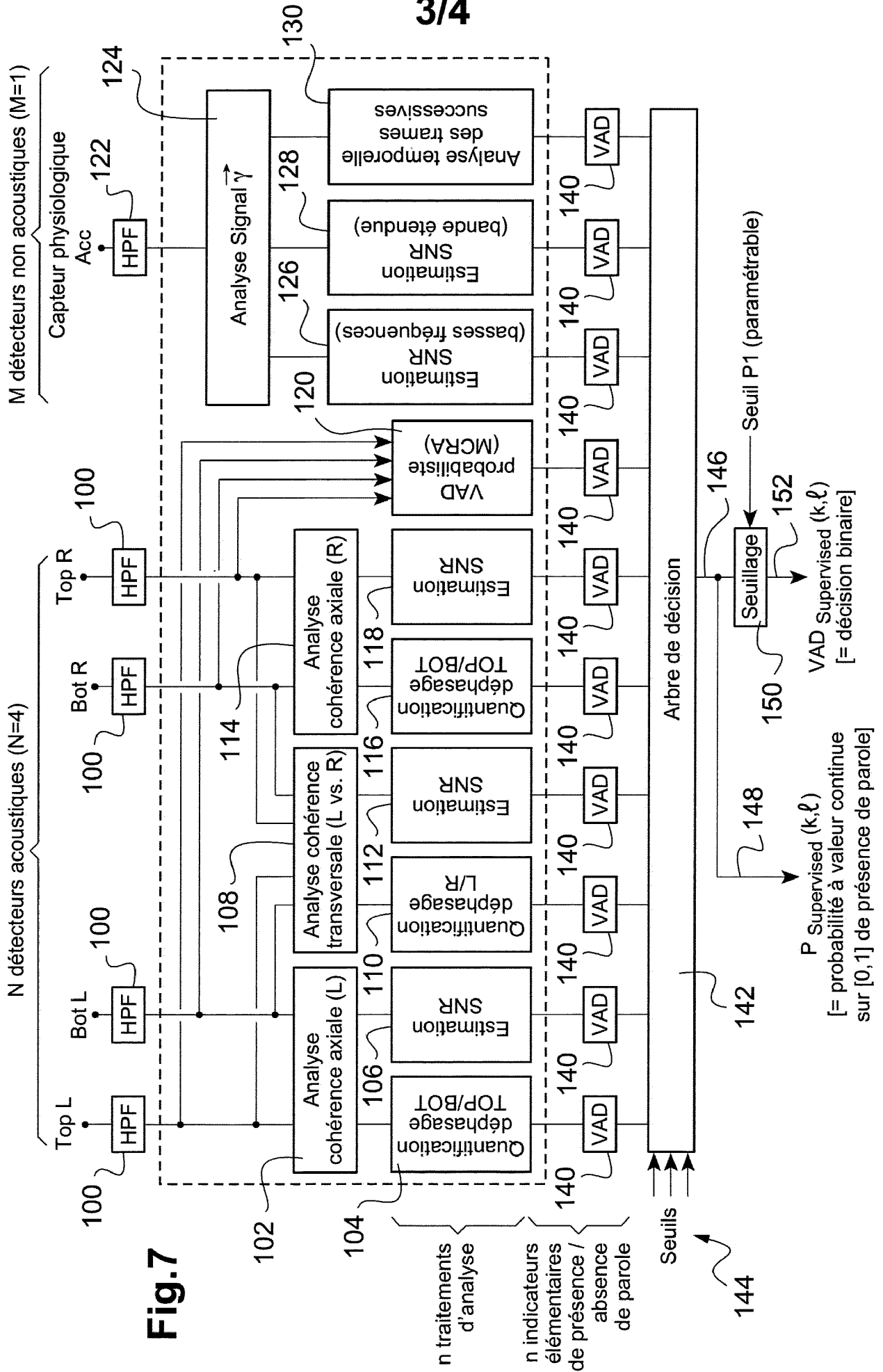


Fig.3

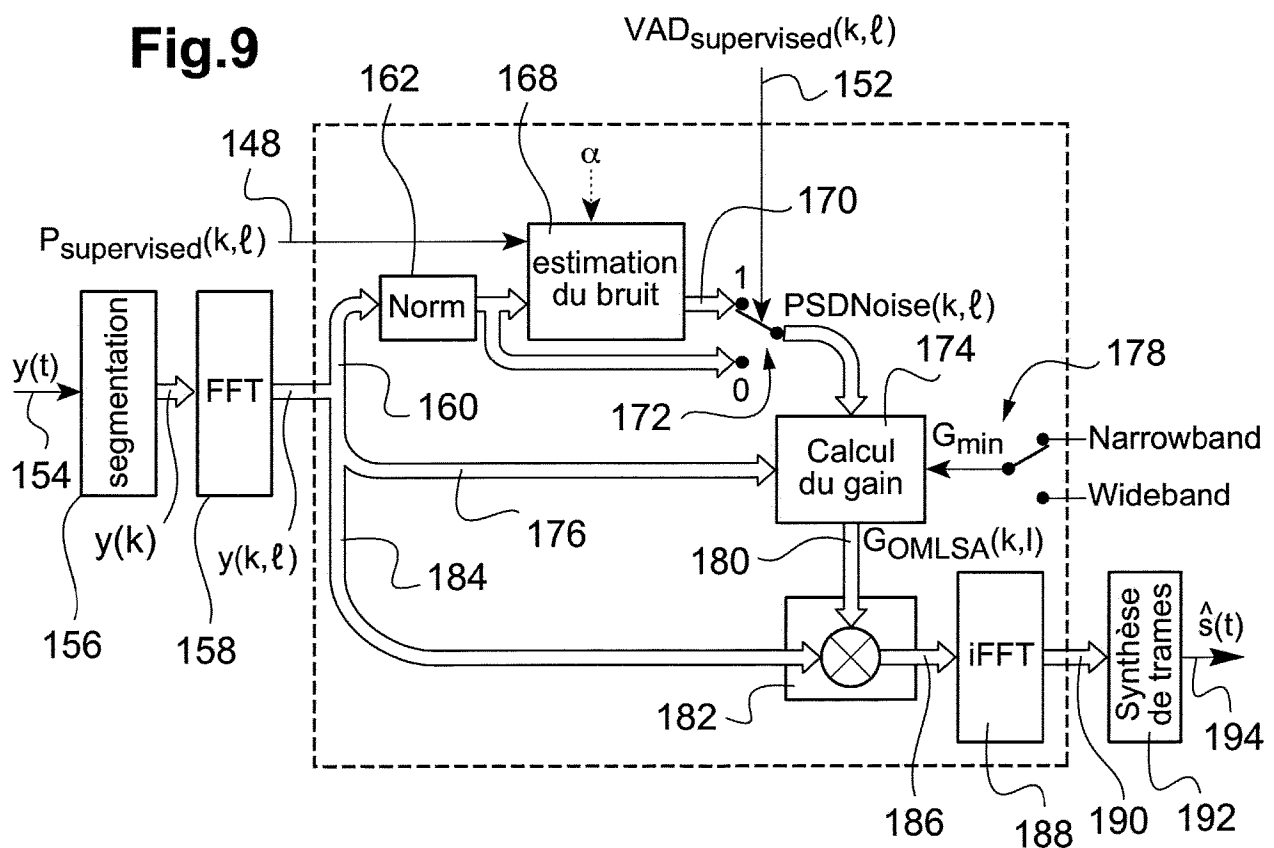
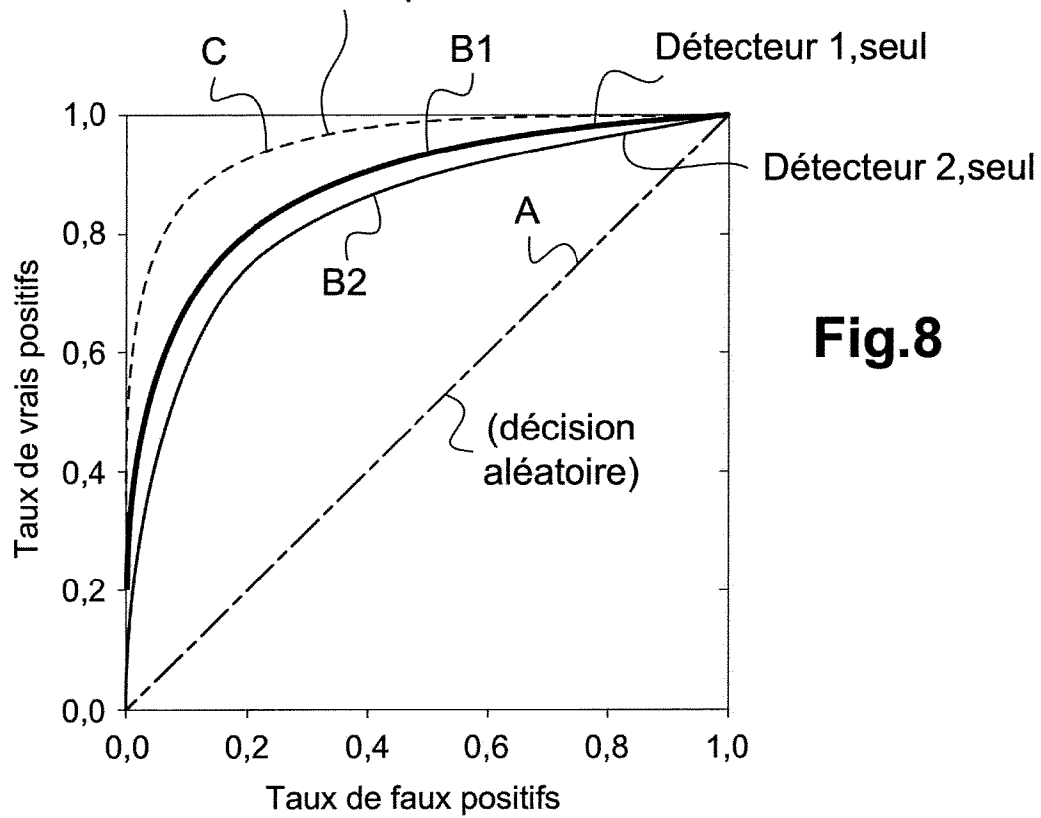
2/4

Fig.4**Fig.5****Fig.6**

3/4



Détecteurs 1 et 2 fusionnés par l'arbre de décision



INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2017/061931

A. CLASSIFICATION OF SUBJECT MATTER
INV. G10L21/0364 H04R1/10 G10L25/78
ADD. H04R3/00

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

B. FIELDS SEARCHED

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

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WON-HO SHIN ET AL: "Speec/non-speech classification using multiple features for robust endpoint detection", ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, 2000. ICASSP '00. PROCEEDING S. 2000 IEEE INTERNATIONAL CONFERENCE ON 5-9 JUNE 2000, PISCATAWAY, NJ, USA, IEEE, vol. 3, 5 June 2000 (2000-06-05), pages 1399-1402, XP010507610, ISBN: 978-0-7803-6293-2 the whole document ----- -/--	1-8



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

11 July 2017

Date of mailing of the international search report

20/07/2017

Name and mailing address of the ISA/

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

Authorized officer

Moscu, Viorel

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2017/061931

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CHOU P A: "OPTIMAL PARTITIONING FOR CLASSIFICATION AND REGRESSION TREES", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE COMPUTER SOCIETY, USA, vol. 13, no. 4, 1 April 1991 (1991-04-01), pages 340-354, XP000214961, ISSN: 0162-8828, DOI: 10.1109/34.88569 the whole document -----	1-8

RAPPORT DE RECHERCHE INTERNATIONALE

Demande internationale n°

PCT/EP2017/061931

A. CLASSEMENT DE L'OBJET DE LA DEMANDE INV. G10L21/0364 H04R1/10 G10L25/78 ADD. H04R3/00		
Selon la classification internationale des brevets (CIB) ou à la fois selon la classification nationale et la CIB		
B. DOMAINES SUR LESQUELS LA RECHERCHE A PORTE Documentation minimale consultée (système de classification suivi des symboles de classement) G10L H04R		
Documentation consultée autre que la documentation minimale dans la mesure où ces documents relèvent des domaines sur lesquels a porté la recherche		
Base de données électronique consultée au cours de la recherche internationale (nom de la base de données, et si cela est réalisable, termes de recherche utilisés) EPO-Internal		
C. DOCUMENTS CONSIDERES COMME PERTINENTS		
Catégorie*	Identification des documents cités, avec, le cas échéant, l'indication des passages pertinents	no. des revendications visées
X	WON-HO SHIN ET AL: "Spec/non-speech classification using multiple features for robust endpoint detection", ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, 2000. ICASSP '00. PROCEEDING S. 2000 IEEE INTERNATIONAL CONFERENCE ON 5-9 JUNE 2000, PISCATAWAY, NJ, USA, IEEE, vol. 3, 5 juin 2000 (2000-06-05), pages 1399-1402, XP010507610, ISBN: 978-0-7803-6293-2 le document en entier ----- -/--	1-8
☒ Voir la suite du cadre C pour la fin de la liste des documents ☐ Les documents de familles de brevets sont indiqués en annexe		
* Catégories spéciales de documents cités:		
"A" document définissant l'état général de la technique, non considéré comme particulièrement pertinent "E" document antérieur, mais publié à la date de dépôt international ou après cette date "L" document pouvant jeter un doute sur une revendication de priorité ou cité pour déterminer la date de publication d'une autre citation ou pour une raison spéciale (telle qu'indiquée) "O" document se référant à une divulgation orale, à un usage, à une exposition ou tous autres moyens "P" document publié avant la date de dépôt international, mais postérieurement à la date de priorité revendiquée	"T" document ultérieur publié après la date de dépôt international ou la date de priorité et n'appartenant pas à l'état de la technique pertinent, mais cité pour comprendre le principe ou la théorie constituant la base de l'invention "X" document particulièrement pertinent; l'invention revendiquée ne peut être considérée comme nouvelle ou comme impliquant une activité inventive par rapport au document considéré isolément "Y" document particulièrement pertinent; l'invention revendiquée ne peut être considérée comme impliquant une activité inventive lorsque le document est associé à un ou plusieurs autres documents de même nature, cette combinaison étant évidente pour une personne du métier "&" document qui fait partie de la même famille de brevets	
Date à laquelle la recherche internationale a été effectivement achevée	Date d'expédition du présent rapport de recherche internationale	
11 juillet 2017	20/07/2017	
Nom et adresse postale de l'administration chargée de la recherche internationale Office Européen des Brevets, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016		Fonctionnaire autorisé Moscu, Viorel

C(suite). DOCUMENTS CONSIDERES COMME PERTINENTS		
Catégorie*	Identification des documents cités, avec, le cas échéant, l'indication des passages pertinents	no. des revendications visées
A	CHOU P A: "OPTIMAL PARTITIONING FOR CLASSIFICATION AND REGRESSION TREES", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE COMPUTER SOCIETY, USA, vol. 13, no. 4, 1 avril 1991 (1991-04-01), pages 340-354, XP000214961, ISSN: 0162-8828, DOI: 10.1109/34.88569 le document en entier -----	1-8