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

(11) **EP 2 680 262 A1**

(12) **DEMANDE DE BREVET EUROPEEN**

(43) Date de publication:
01.01.2014 Bulletin 2014/01

(51) Int Cl.:
G10L 21/0208 *(2013.01)* ***G10L 21/0216*** *(2013.01)*

(21) Numéro de dépôt: **13171948.6**

(22) Date de dépôt: **14.06.2013**

(84) Etats contractants désignés:
AL 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 RS SE SI SK SM TR
Etats d'extension désignés:
BA ME

(30) Priorité: **26.06.2012 FR 1256049**

(71) Demandeur: **Parrot**
75010 Paris (FR)

(72) Inventeurs:
- **Fox, Charles**
75014 Paris (FR)
- **Vitte, Guillaume**
75003 Paris (FR)
- **Charbit, Maurice**
94800 Villejuif (FR)
- **Prado, Jacques**
35120 Cherrueix (FR)

(74) Mandataire: **Dupuis-Latour, Dominique**
Bardehle Pagenberg
10, boulevard Haussmann
75009 Paris (FR)

(54) **Procédé de débruitage d'un signal acoustique pour un dispositif audio multi-microphone opérant dans un milieu bruité**

(57) Ce procédé comporte des étapes de : a) partition (10, 16) du spectre du signal bruité en une partie HF et une partie BF ; b) traitements de débruitage opérés de façon différenciée pour chacune des deux parties du spectre, avec pour la partie HF un débruitage par prédiction du signal utile d'un capteur sur l'autre entre capteurs d'un premier sous-réseau (R_1), au moyen d'un premier estimateur (14) à algorithme adaptatif, et pour la partie BF un débruitage par prédiction du bruit d'un capteur sur l'autre entre capteurs d'un second sous-réseau (R_2), au moyen d'un second estimateur (18) à algorithme adaptatif ; c) reconstruction du spectre par combinaison (22) des signaux délivrés après les traitements de débruitage respectifs des deux parties du spectre ; et d) réduction sélective du bruit (24) par un traitement de gain à amplitude log-spectrale modifié optimisé, OM-LSA.

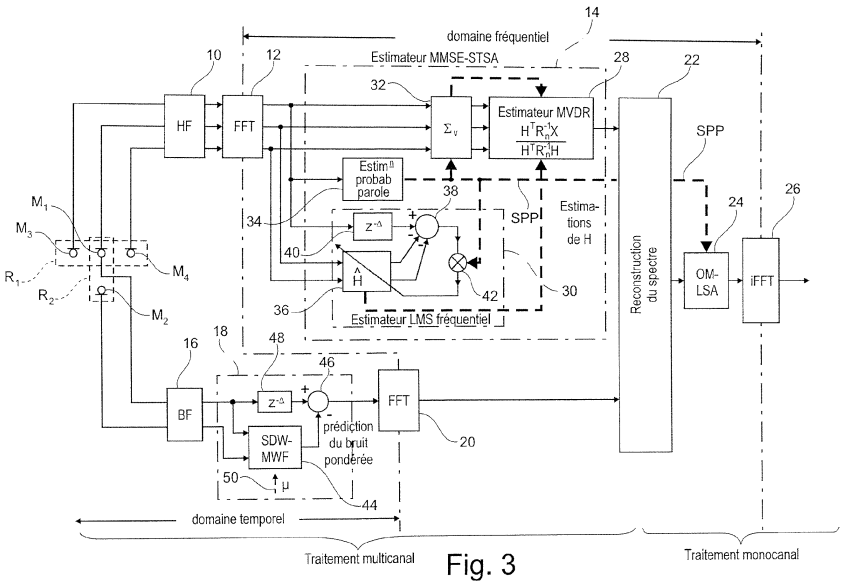

Fig. 3

EP 2 680 262 A1

Description

[0001] L'invention concerne le traitement de la parole en milieu bruité.

[0002] Elle concerne notamment le traitement des signaux de parole captés par des dispositifs de téléphonie de type "mains libres" destinés à être utilisés dans un environnement bruité.

[0003] Ces appareils comportent un ou plusieurs microphones ("micros") sensibles, captant non seulement la voix de l'utilisateur, mais également le bruit environnant, bruit qui constitue un élément perturbateur pouvant aller dans certains cas jusqu'à rendre inintelligibles les paroles du locuteur. Il en est de même si l'on veut mettre en oeuvre des techniques de reconnaissance vocale, car il est très difficile d'opérer une reconnaissance de forme sur des mots noyés dans un niveau de bruit élevé.

[0004] Cette difficulté liée aux bruits environnants est particulièrement contraignante dans le cas des dispositifs "mains libres" pour véhicules automobiles, qu'il s'agisse d'équipements incorporés au véhicule ou bien d'accessoires en forme de boitier amovible intégrant tous les composants et fonctions de traitement du signal pour la communication téléphonique.

[0005] En effet, dans cette application, la distance importante entre le micro (placé au niveau de la planche de bord ou dans un angle du pavillon de l'habitacle) et le locuteur (dont l'éloignement est contraint par la position de conduite) entraine la captation d'un niveau de bruit relativement élevé, qui rend difficile l'extraction du signal utile noyé dans le bruit. De plus, le milieu très bruité typique de l'environnement automobile présente des caractéristiques spectrales qui évoluent de manière imprévisible en fonction des conditions de conduite : passage sur des chaussées déformées ou pavées, autoradio en fonctionnement, etc.

[0006] Des difficultés comparables se présentent lorsque le dispositif est un casque audio de type micro/casque combiné utilisé pour des fonctions de communication telles que des fonctions de téléphonie "mains libres", en complément de l'écoute d'une source audio (musique par exemple) provenant d'un appareil sur lequel est branché le casque.

[0007] Dans ce cas, il s'agit d'assurer une intelligibilité suffisante du signal capté par le micro, c'est-à-dire du signal de parole du locuteur proche (le porteur du casque). Or le casque peut être utilisé dans un environnement bruyant (métro, rue passante, train, etc.), de sorte que le micro captera non seulement la parole du porteur du casque, mais également les bruits parasites environnants. Le porteur est protégé de ce bruit par le casque, notamment s'il s'agit d'un modèle à écouteurs fermés isolant l'oreille de l'extérieur, et encore plus si le casque est pourvu d'un "contrôle actif de bruit". En revanche, le locuteur distant (celui se trouvant à l'autre bout du canal de communication) souffrira des bruits parasites captés par le micro et venant se superposer et interférer avec le signal de parole du locuteur proche (le porteur du casque). En particulier, certains formants de la parole essentiels à la compréhension de la voix sont souvent noyés dans des composantes de bruit couramment rencontrées dans les environnements habituels.

[0008] L'invention concerne plus particulièrement les techniques de débruitage mettant en oeuvre un réseau de plusieurs micros, en combinant de façon judicieuse les signaux captés simultanément par ces micros pour discriminer les composantes utiles de parole d'avec les composantes parasites de bruit.

[0009] Une technique classique consiste à placer et orienter l'un des micros pour qu'il capte principalement la voix du locuteur, tandis que l'autre est disposé de manière à capter une composante de bruit plus importante que le micro principal. La comparaison des signaux captés permet d'extraire la voix du bruit ambiant par analyse de cohérence spatiale des deux signaux, avec des moyens logiciels relativement simples.

[0010] Le US 2008/0280653 A1 décrit une telle configuration, où l'un des micros (celui qui capte principalement la voix) est celui d'une oreillette sans fil portée par le conducteur du véhicule, tandis que l'autre (celui qui capte principalement le bruit) est celui de l'appareil téléphonique, placé à distance dans l'habitacle du véhicule, par exemple accroché au tableau de bord.

[0011] Cette technique présente cependant l'inconvénient de nécessiter deux micros distants, l'efficacité étant d'autant plus élevée que les deux micros sont éloignés. De ce fait, cette technique n'est pas applicable à un dispositif dans lequel les deux micros sont rapprochés, par exemple deux micros incorporés à la façade d'un autoradio de véhicule automobile, ou deux micros qui seraient disposés sur l'une des coques d'un écouteur de casque audio.

[0012] Une autre technique encore, dite *beamforming,* consiste à créer par des moyens logiciels une directivité qui améliore le rapport signal/bruit du réseau ou "antenne" de micros. Le US 2007/0165879 A1 décrit une telle technique, appliquée à une paire de micros non directionnels placés dos à dos. Un filtrage adaptatif des signaux captés permet de dériver en sortie un signal dans lequel la composante de voix a été renforcée.

[0013] Toutefois, on estime qu'une méthode de débruitage multi-capteurs ne fournit de bons résultats qu'à condition de disposer d'un réseau d'au moins huit micros, les performances étant extrêmement limitées lorsque seulement deux micros sont utilisés.

[0014] Les EP 2 293 594 A1 et EP 2 309 499 A1 (Parrot) décrivent d'autres techniques, également basées sur l'hypothèse que le signal utile et/ou les bruits parasites présentent une certaine directivité, qui combinent les signaux issus des différents micros de manière à améliorer le rapport signal/bruit en fonction de ces conditions de directivité. Ces techniques de débruitage reposent sur l'hypothèse que la parole présente généralement une cohérence spatiale

supérieure au bruit et que, par ailleurs, la direction d'incidence de la parole est généralement bien définie et peut être supposée connue (dans le cas d'un véhicule automobile, elle est définie par la position du conducteur, vers lequel sont tournés les micros). Cette hypothèse prend cependant mal en compte l'effet de réverbération typique de l'habitacle d'une voiture, où les réflexions puissantes et nombreuses rendent difficile le calcul d'une direction d'arrivée. Elles peuvent être également mises en défaut par des bruits présentant une certaine directivité, tels que coups de klaxon, passage d'un scooter, dépassement par une voiture, etc.

[0015] Un autre procédé encore est décrit dans l'article de I. McCowan et S. Sridharan, "Adaptive Parameter Compensation for Robust Hands-free Speech Recognition using a Dual-Beamforming Microphone Array", Proceedings on 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing, May 2001.

[0016] De façon générale, ces techniques basées sur des hypothèses de directivité présentent toutes des performances limitées à l'encontre des composantes de bruit situées dans la région des fréquences les plus basses - là où, précisément, le bruit peut se trouver concentré à un niveau d'énergie relativement élevé.

[0017] En effet, la directivité est d'autant plus marquée que la fréquence est élevée, de sorte que ce critère devient peu discriminant pour les fréquences les plus basses. En fait, pour rester suffisamment efficace, il est nécessaire d'écarter beaucoup les micros, par exemple de 15 à 20 cm, voire même plus en fonction des performances souhaitées, de manière à décorréler suffisamment les bruits captés par ces micros.

[0018] Par voie de conséquence, il n'est pas possible d'incorporer un tel réseau de micros par exemple au boitier d'un autoradio de véhicule automobile ou à un boitier de "kit mains libres" autonome placé dans le véhicule, encore moins sur des coques d'écouteurs d'un casque audio.

[0019] Le problème de l'invention est, dans un tel contexte, de pouvoir disposer d'une technique de réduction de bruit efficace permettant de délivrer au locuteur distant un signal vocal représentatif de la parole émise par le locuteur proche (conducteur du véhicule ou porteur du casque), en débarrassant ce signal des composantes parasites de bruit extérieur présentes dans l'environnement de ce locuteur proche, technique qui :

- présente des performances accrues dans le bas du spectre des fréquences, là où sont le plus souvent concentrées les composantes de bruit parasite les plus gênantes, notamment du point de vue du masquage du signal de parole ;
- ne requière pour sa mise en oeuvre qu'un nombre réduit de micros (typiquement, pas plus de trois à cinq micros) ; et
- avec une configuration géométrique suffisamment ramassée du réseau de micros (typiquement avec un écartement entre micros de quelques centimètres seulement), pour permettre notamment son intégration à des produits compacts de type "tout-en-un".

[0020] Le point de départ de l'invention réside dans l'analyse du champ de bruit typique dans l'habitacle d'un véhicule automobile, qui conduit aux observations suivantes :

- le bruit dans l'habitacle est spatialement cohérent dans les basses fréquences (au-dessous de 1000 Hz environ) ;
- il perd en cohérence dans les hautes fréquences (au-dessus de 1000 Hz) ; et
- selon le type de micro utilisé, unidirectionnel ou omnidirectionnel, la cohérence spatiale est modifiée.

[0021] Ces observations, qui seront précisées et justifiées plus loin, conduisent à proposer une stratégie de débruitage hybride, mettant en oeuvre en basse fréquence (BF) et en haute fréquence (HF) deux algorithmes différents, exploitant la cohérence ou la non-cohérence des composantes de bruit selon la partie du spectre considérée :

- la forte cohérence des bruits en BF permet d'envisager un algorithme exploitant une prédiction du bruit d'un micro sur l'autre, ce qui est possible car on peut observer des périodes de silence du locuteur, avec absence de signal utile et présence exclusive du bruit ;
- en revanche, en HF le bruit est faiblement cohérent et il est difficilement prédictible, sauf à prévoir un nombre élevé de micros (ce qui n'est pas souhaité) ou à rapprocher les micros pour rendre les bruits plus cohérents (mais l'on n'obtiendra jamais de grande cohérence dans cette bande, sauf à confondre les micros : les signaux captés seraient alors les mêmes, et l'on n'aurait aucune information spatiale). Pour cette partie HF, on utilisera alors un algorithme exploitant le caractère prédictible du signal utile d'un micro sur l'autre (et non plus une prédiction du bruit), ce qui est par hypothèse possible car on sait que ce signal utile est produit par une source ponctuelle (la bouche du locuteur).

[0022] Plus précisément, l'invention propose un procédé de débruitage d'un signal acoustique bruité pour un dispositif audio multi-microphone du type général divulgué par l'article précité de McCowan and S. Sridharan, où le dispositif comprend un réseau de capteurs formé d'une pluralité de capteurs microphoniques disposés selon une configuration prédéterminée et aptes à recueillir le signal bruité, les capteurs étant regroupés en deux sous-réseaux, avec un premier sous-réseau apte à recueillir une partie HF du spectre, et un second sous-réseau de capteurs apte à recueillir une partie BF du spectre distincte de la partie HF.

[0023] Ce procédé comporte des étapes de :

a) partition du spectre du signal bruité entre ladite partie HF et ladite partie BF, par filtrage respectivement au-delà et en deçà d'une fréquence pivot prédéterminée,
b) débruitage de chacune des deux parties du spectre avec mise en oeuvre d'un estimateur à algorithme adaptatif ; et
c) reconstruction du spectre par combinaison des signaux délivrés après débruitage des deux parties du spectre aux étapes b1) et b2),

[0024] De façon caractéristique de l'invention, l'étape b) de débruitage est opérée par des traitements distincts pour chacune des deux parties du spectre, avec :

b1) pour la partie HF, un débruitage exploitant le caractère prédictible du signal utile d'un capteur sur l'autre entre capteurs du premier sous-réseau, au moyen d'un premier estimateur (14) à algorithme adaptatif, et
b2) pour la partie BF, un débruitage par prédiction du bruit d'un capteur sur l'autre entre capteurs du second sous-réseau, au moyen d'un second estimateur (18) à algorithme adaptatif.

[0025] En ce qui concerne la géométrie du réseau de capteurs, le premier sous-réseau de capteurs apte à recueillir la partie HF du spectre peut notamment comprendre un réseau linaire d'au moins deux capteurs alignés perpendiculairement à la direction de la source de parole, et le second sous-réseau de capteurs apte à recueillir la partie BF du spectre peut comprendre un réseau linaire d'au moins deux capteurs alignés parallèlement à la direction de la source de parole.

[0026] Les capteurs du premier sous-réseau de capteurs sont avantageusement des capteurs unidirectionnels, orientés dans la direction de la source de parole.

[0027] Le traitement de débruitage de la partie HF du spectre à l'étape b1) peut être opéré de façon différenciée pour une bande inférieure et une bande supérieure de cette partie HF, avec sélection de capteurs différents parmi les capteurs du premier sous-réseau, la distance entre les capteurs sélectionnés pour le débruitage de la bande supérieure étant plus réduite que la distance des capteurs sélectionnés pour le débruitage de la bande inférieure.

[0028] Le traitement de débruitage prévoit de préférence, après l'étape c) de reconstruction du spectre, une étape de :

d) réduction sélective du bruit par un traitement de type gain à amplitude log-spectrale modifié optimisé, OM-LSA, à partir du signal reconstruit produit à l'étape c) et d'une probabilité de présence de parole.

[0029] En ce qui concerne le débruitage de la partie HF du spectre, l'étape b1), exploitant le caractère prédictible du signal utile d'un capteur sur l'autre, peut être opérée dans le domaine fréquentiel, en particulier par :

b11) estimation d'une probabilité de présence de parole dans le signal bruité recueilli ;
b12) estimation d'une matrice spectrale de covariance des bruits recueillis par les capteurs du premier sous-réseau, cette estimation étant modulée par la probabilité de présence de parole ;
b13) estimation de la fonction de transfert des canaux acoustiques entre la source de parole et au moins certains des capteurs du premier sous-réseau, cette estimation étant opérée par rapport à une référence de signal utile constituée par le signal recueilli par l'un des capteurs du premier sous-réseau, et étant en outre modulée par la probabilité de présence de parole ; et
b14) calcul, notamment par un estimateur de type *beamforming* à réponse sans distorsion à variance minimale, MVDR, d'un projecteur linéaire optimal donnant un signal combiné débruité unique à partir des signaux recueillis par au moins certains des capteurs du premier sous-réseau, de la matrice spectrale de covariance estimée à l'étape b12), et des fonctions de transfert estimées à l'étape b13).

[0030] L'étape b13) d'estimation de la fonction de transfert des canaux acoustiques peut notamment être mise en oeuvre par un filtre adaptatif à prédiction linéaire de type moindres carrés moyens, LMS, avec modulation par la probabilité de présence de parole, notamment une modulation par variation du pas d'itération du filtre adaptatif LMS.

[0031] Pour le débruitage de la partie BF à l'étape b2), la prédiction du bruit d'un capteur sur l'autre peut être opérée dans le domaine f temporel, en particulier par un filtre de type filtre de Wiener multicanal avec pondération par la distorsion de la parole, SDW-MWF, notamment un filtre SDW-MWF estimé de manière adaptative par un algorithme de descente de gradient.

[0032] On va maintenant décrire un exemple de mise en oeuvre du dispositif de l'invention, en référence aux dessins annexés où les mêmes références numériques désignent d'une figure à l'autre des éléments identiques ou fonctionnellement semblables.

La Figure 1 illustre de façon schématique un exemple de réseau de micros, comprenant quatre micros utilisables de façon sélective pour la mise en oeuvre de l'invention.

Les Figures 2a et 2b sont des caractéristiques, respectivement pour un micro omnidirectionnel et pour un micro unidirectionnel, montrant les variations, en fonction de la fréquence, de la corrélation (fonction de cohérence quadratique) entre deux micros pour un champ de bruit diffus, ceci pour plusieurs valeurs d'écartement entre ces deux micros.

La Figure 3 est un schéma d'ensemble, sous forme de blocs fonctionnels, montrant les différents traitements selon l'invention pour le débruitage des signaux recueillis par le réseau de micros de la Figure 1.

La Figure 4 est une représentation schématique par blocs fonctionnels, généralisée à un nombre de micros supérieur à deux, d'un filtre adaptatif pour l'estimation de la fonction de transfert d'un canal acoustique, utilisable pour le traitement de débruitage de la partie BF du spectre dans le traitement d'ensemble de la Figure 3.

[0033] On va maintenant décrire en détail un exemple de technique de débruitage mettant en oeuvre les enseignements de l'invention.

Configuration du réseau de capteurs microphoniques

[0034] On considèrera, comme illustré Figure 1, un réseau R de capteurs microphoniques M_1 ... M_4, chaque capteur pouvant être assimilé à un micro unique captant une version bruitée d'un signal de parole émis par une source de signal utile (locuteur) de direction d'incidence Δ.

[0035] Chaque micro capte donc une composante du signal utile (le signal de parole) et une composante du bruit parasite environnant, sous toutes ses formes (directif ou diffus, stationnaire ou évoluant de manière imprévisible, etc.).

[0036] Le réseau R est configuré en deux sous-réseaux R_1 et R_2 dédiés respectivement à la captation et au traitement des signaux dans la partie supérieure (ci-après "haute fréquence", HF) du spectre et dans la partie inférieure (ci-après "basse fréquence", BF) de ce même spectre.

[0037] Le sous-réseau R_1 dédié à la partie HF du spectre est constitué des trois micros M_1, M_3, M_4 qui sont alignés perpendiculairement à la direction d'incidence Δ, avec un écartement respectif de d = 2 cm dans l'exemple illustré. Ces micros sont de préférence des micros unidirectionnels dont le lobe principal est orienté dans la direction Δ du locuteur.

[0038] Le sous-réseau R_2 dédié à la partie BF du spectre est constitué des deux micros M_1 et M_2, alignés parallèlement à la direction Δet écartés de d = 3 cm dans l'exemple illustré. On notera que le micro M_1, qui appartient aux deux sous-réseaux R_1 et R_2, est mutualisé, ce qui permet de réduire le nombre total de micros du réseau. Cette mutualisation est avantageuse mais elle n'est toutefois pas nécessaire. D'autre part, on a illustré une configuration en forme de "L" où le micro mutualisé est le micro M_1, mais cette configuration n'est pas restrictive, le micro mutualisé pouvant être par exemple le micro M_3, donnant à l'ensemble du réseau une configuration en forme de "T".

[0039] Par ailleurs, le micro M_2 du réseau BF peut être un micro omnidirectionnel, dans la mesure où la directivité est beaucoup moins marquée en BF qu'en HF.

[0040] Enfin, la configuration illustrée montrant deux sous-réseaux R_1 + R_2 comprenant 3 + 2 micros (soit un total de 4 micros compte tenu de la mutualisation de l'un des micros) n'est pas limitative. La configuration minimale est une configuration à 2 + 2 micros (soit un minimum de 3 micros si l'un d'entre eux est mutualisé). Inversement, il est possible d'augmenter le nombre de micros, avec des configurations à 4 + 2 micros, 4 + 3 micros, etc.

[0041] L'augmentation du nombre de micros permet, notamment dans les hautes fréquences, de sélectionner des configurations de micros différentes selon les parties du spectre HF traitées.

[0042] Ainsi, dans l'exemple illustré, si l'on opère en téléphonie *wideband* avec une plage de fréquences allant jusqu'à 8000 Hz (au lieu de 4000 Hz), pour la bande inférieure (1000 à 4000 Hz) de la partie HF du spectre on choisira les deux micros extrêmes {M_1, M_4} éloignés entre eux de d = 4 cm, tandis que pour la bande supérieure (4000 à 8000 Hz) de cette même partie HF on utilisera un couple de deux micros voisins {M_1, M_3} ou {M_3, M_4}, ou bien les trois micros {M_1, M_3, M_4} ensemble, ces micros étant espacés chacun de d = 2 cm seulement: on bénéficie ainsi dans la bande inférieure du spectre HF de l'écartement maximum des micros, ce qui maximise la décorrélation des bruits captés, tout en évitant dans la bande supérieure un repliement des hautes fréquences du signal à restituer ; un tel repliement apparaitrait sinon du fait d'une fréquence d'échantillonnage spatiale trop faible, dans la mesure où il faut que le retard de phase maximal d'un signal capté par un micro puis par l'autre soit inférieur à la période d'échantillonnage du convertisseur de numérisation des signaux. On va maintenant exposer, en référence aux Figures 2a et 2b, la manière de choisir la fréquence pivot entre les deux parties BF et HF du spectre, et le choix préférentiel du type de micro unidirectionnel/omnidirectionnel selon la partie du spectre à traiter, HF ou BF.

[0043] Ces Figures 2a et 2b illustrent, respectivement pour un micro omnidirectionnel et pour un micro unidirectionnel, des caractéristiques donnant, en fonction de la fréquence, la valeur de la fonction de corrélation entre deux micros, pour plusieurs valeurs d d'écartement entre ces micros.

[0044] La fonction de corrélation entre deux micros éloignés d'une distance d, pour un modèle champ de bruit diffus,

est une fonction globalement décroissante de la distance entre les micros. Cette fonction de corrélation est représentée par la cohérence quadratique moyenne *MSC* (*Mean Squared Coherence*), qui varie entre 1 (les deux signaux sont parfaitement cohérents, ils ne diffèrent que d'un filtre linéaire) et 0 (signaux totalement décorrélés). Dans le cas d'un micro omnidirectionnel, cette cohérence peut être modélisée en fonction de la fréquence par la fonction :

$$MSC(f) = \left|\frac{\sin(2\pi f\tau)}{2\pi f\tau}\right|^2$$

f étant la fréquence considérée et τ étant le retard de propagation entre les micros soit $\tau = d/c$, où *d* est la distance entre les micros et *c* la vitesse du son.

[0045] Cette courbe modélisée a été illustrée sur la Figure 2a, les figures 2a et 2b montant également la fonction de cohérence *MSC* réellement mesurée pour les deux types de micros et pour diverses valeurs de distances *d.*

[0046] Si l'on considère que l'on est en présence de signaux effectivement cohérents lorsque la valeur de *MSC* > 0,9, le bruit pourra être considéré comme étant cohérent lorsque l'on se trouve au-dessous d'une fréquence f_0 telle que :

$$f_0 = \frac{0{,}787c}{2\pi d}.$$

[0047] Ceci donne une fréquence pivot f_0 d'environ 1000 Hz pour des micros écartés de d = 4 cm (distance entre les micros M_1 et M_4 de l'exemple de réseau de la Figure 1).

[0048] Dans le présent exemple, correspondant notamment au réseau de micros ayant les dimensions indiquées plus haut, on choisira ainsi une fréquence pivot f_0 = 1000 Hz au-dessous de laquelle (partie BF) on considèrera que le bruit est cohérent, ce qui permet d'envisager un algorithme basé sur une prédiction de ce bruit d'un micro sur l'autre (prédiction opérée pendant les périodes de silence du locuteur, où seul le bruit est présent).

[0049] De préférence, on utilisera pour cette partie BF, des micros unidirectionnels, car comme on peut le voir en comparant les Figures 2a et 2b la variation de la fonction de cohérence est beaucoup plus abrupte dans ce cas qu'avec un micro omnidirectionnel.

[0050] Dans la partie HF du spectre, où le bruit est faiblement cohérent, il n'est plus possible de prédire ce bruit de façon satisfaisante ; on mettra alors en oeuvre un autre algorithme, exploitant le caractère prédictible du signal utile (et non plus du bruit) d'un micro sur l'autre.

[0051] On notera enfin que le choix de la fréquence pivot (f_0 = 1000 Hz pour d = 2 cm) dépend aussi de l'écartement entre micros, un écartement plus grand correspondant à une fréquence pivot plus faible, et *vice versa.*

Traitement de débruitage : description d'un mode préférentiel

[0052] On va maintenant décrire, en référence à la Figure 3, un mode de mise en oeuvre préférentiel de débruitage des signaux recueillis par le réseau de micros de la Figure 1, de façon bien entendu non limitative.

[0053] Comme expliqué plus haut, des traitements différents sont opérés pour le haut du spectre (hautes fréquences, HF) et pour le bas du spectre (basses fréquences, BF).

[0054] Pour le haut du spectre, un filtre passe-haut HF 10 reçoit les signaux des micros M_1, M_3 et M_4 du sous-réseau R_1, utilisés conjointement. Ces signaux font d'abord l'objet d'une transformée rapide de Fourier FFT (bloc 12), puis d'un traitement, dans le domaine fréquentiel, par un algorithme (bloc 14) exploitant le caractère prédictible du signal utile d'un micro sur l'autre, dans cet exemple un estimateur de type MMSE-STSA (*Minimum Mean-Squared Error Short-Time Spectral Amplitude*), qui sera décrit en détail plus bas.

[0055] Pour le bas du spectre, un filtre passe-bas BF 16 reçoit en entrée les signaux captés par les micros M_1 et M_2 du sous-réseau R_2. Ces signaux font l'objet d'un traitement de débruitage (bloc 18) opéré dans le domaine temporel par un algorithme exploitant une prédiction du bruit d'un micro sur l'autre pendant les périodes de silence du locuteur. Dans cet exemple, on utilise un algorithme de type SDW-MWF (*Speech Distorsion Weighted Multichannel Wiener Filter*), qui sera décrit plus en détail par la suite. Le signal débruité résultant fait ensuite l'objet d'une transformée rapide de Fourier FFT (bloc 20).

[0056] On dispose ainsi, à partir de deux traitements multicanal, de deux signaux monocanal résultants, l'un pour la partie HF issue du bloc 14, l'autre pour la partie BF issue du bloc 18 après passage dans le domaine fréquentiel par le bloc 20.

[0057] Ces deux signaux résultants débruités sont combinés (bloc 22) de manière à opérer une reconstruction du spectre complet, HF + BF.

[0058] Très avantageusement, un traitement (monocanal) supplémentaire de débruitage sélectif (bloc 24) est opéré

sur le signal reconstruit correspondant. Le signal issu de ce traitement fait enfin l'objet d'une transformée de Fourier rapide inverse iFFT (bloc 26) pour repasser dans le domaine temporel.

[0059] Plus précisément, ce traitement de débruitage sélectif final consiste à appliquer un gain variable propre à chaque bande de fréquence, ce débruitage étant également modulé par une probabilité de présence de parole.

[0060] On peut avantageusement utiliser pour le débruitage du bloc 24 une méthode de type OM/LSA (*Optimally Modified - Log Spectral Amplitude*) telle que celle décrite par :

[1] I. Cohen, "Optimal Speech Enhancement under Signal Presence Uncertainty Using Log-Spectral Amplitude Estimator", Signal Processing Letters, IEEE, Vol. 9, No 4, pp. 113-116, Apr. 2002.

Essentiellement, l'application d'un gain nommé "gain LSA" (*Log-Spectral Amplitude*) permet de minimiser la distance quadratique moyenne entre le logarithme de l'amplitude du signal estimé et le logarithme de l'amplitude du signal de parole originel. Ce second critère se montre supérieur au premier car la distance choisie est en meilleure adéquation avec le comportement de l'oreille humaine et donne donc qualitativement de meilleurs résultats.

Dans tous les cas, il s'agit de diminuer l'énergie des composantes fréquentielles très parasitées en leur appliquant un gain faible, tout en laissant intactes (par l'application d'un gain égal à 1) celles qui le sont peu ou pas du tout.

L'algorithme "OM-LSA" (*Optimally-Modified LSA*) améliore le calcul du gain LSA à appliquer en le pondérant par une probabilité conditionnelle de présence de parole SPP (*Speech Presence Probability*), qui intervient à deux niveaux :

- pour l'estimation de l'énergie du bruit : la probabilité module le facteur d'oubli dans le sens d'une mise à jour plus rapide de l'estimation du bruit sur le signal bruité lorsque la probabilité de présence de parole est faible ;
- pour le calcul du gain final : la réduction de bruit appliquée est d'autant plus importante (c'est-à-dire que le gain appliqué est d'autant plus faible) que la probabilité de présence de parole est faible.

[0061] La probabilité de présence de parole SPP est un paramètre pouvant prendre plusieurs valeurs différentes comprises entre 0 et 100 %. Ce paramètre est calculé selon une technique en elle-même connue, dont des exemples sont notamment exposés dans :

[2] I. Cohen et B. Berdugo, "Two-Channel Signal Detection and Speech Enhancement Based on the Transient Beam-to-Reference Ratio", IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2003, Hong-Kong, pp. 233-236, Apr. 2003.

[0062] On pourra également se référer au WO 2007/099222 A1 (Parrot), qui décrit une technique de débruitage mettant en oeuvre un calcul de probabilité de présence de parole.

Algorithme MMSE-STSA de débruitage HF (bloc 14)

[0063] On va décrire un exemple de traitement de débruitage appliqué à la partie HF du spectre, par un estimateur MMSE-STSA opérant dans le domaine fréquentiel.

[0064] Cette mise en oeuvre particulière n'est bien entendu pas limitative, d'autres techniques de débruitage pouvant être envisagées, dès lors qu'elles sont basées sur le caractère prédictible du signal utile d'un micro sur l'autre. En outre, ce débruitage HF n'est pas nécessairement opéré dans le domaine fréquentiel, il peut également être opéré dans le domaine temporel, par des moyens équivalents.

[0065] La technique proposée consiste à rechercher un "projecteur" linéaire optimal pour chaque fréquence, c'est-à-dire un opérateur correspondant à une transformation d'une pluralité de signaux (ceux recueillis concurremment par les divers micros du sous-réseau R_1) en un signal unique monocanal.

[0066] Cette projection, estimée par le bloc 28, est une projection linéaire "optimale" en ce sens que l'on cherche à ce que la composante de bruit résiduel sur le signal monocanal délivré en sortie soit minimisée et que la composante utile de parole soit la moins déformée possible.

[0067] Cette optimisation implique de rechercher pour chaque fréquence un vecteur A tel que :

- la projection $\boldsymbol{A^T X}$ contienne le moins de bruit possible, c'est-à-dire que la puissance du bruit résiduel, qui vaut $E[\boldsymbol{A^T V V^T A}] = \boldsymbol{A^T R_n A}$, soit minimisée, et
- la voix du locuteur ne soit pas déformée, ce qui se traduit par la contrainte $\boldsymbol{A^T H}=1$, où $\boldsymbol{R_n}$ est la matrice de corrélation entre les micros, pour chaque fréquence, et $\boldsymbol{H}$ est le canal acoustique considéré.

[0068] Ce problème est un problème d'optimisation sous contrainte, à savoir la recherche de $min(\boldsymbol{A^T R_n A})$ sous la contrainte $\boldsymbol{A^T H}=1$.

[0069] Il peut être résolu en utilisant la méthode des multiplieurs de Lagrange, qui conduit à la solution : $A^T = \frac{H^T R_n^{-1}}{H^T R_n^{-1} H}$.

[0070] Dans le cas où les fonctions de transfert H correspondent à un retard pur, on reconnait la formule du *beamforming* MVDR (*Minimum Variance Distorsionless Response*), aussi appelé *beamforming* de Capon. On notera que la puissance de bruit résiduel vaut, après projection $\frac{1}{H^T R_n^{-1} H}$.

[0071] De plus, si l'on considère des estimateurs de type MMSE (*Minimum Mean-Squared Error*) sur l'amplitude et la phase du signal à chaque fréquence, on constate que ces estimateurs s'écrivent comme un *beamforming* de Capon suivi d'un traitement monocanal de débruitage sélectif, comme cela a été exposé par :

[3] R. C. Hendriks et al., On optimal multichannel mean-squared error estimators for speech enhancement, IEEE Signal Processing Letters, vol. 16, no. 10, 2009.

[0072] Le traitement de débruitage sélectif du bruit, appliqué au signal monocanal résultant du traitement de *beamforming,* est avantageusement le traitement de type OM-LSA décrit plus haut, opéré par le bloc 24 sur le spectre complet après synthèse en 22.

[0073] La matrice interspectrale des bruits est estimée récursivement (bloc 32), en utilisant la probabilité de présence de parole *SPP* (bloc 34, voir plus haut) :

$$\Sigma_{bb}(t) = \alpha \Sigma_{bb}(t-1) + (1-\alpha)\mathbf{X}(t)\mathbf{X}(t)^T$$
$$\alpha = \alpha_0 + (1-\alpha_0)SPP$$

α_0 étant un facteur d'oubli.

[0074] En ce qui concerne l'estimateur MVDR (bloc 28), sa mise en oeuvre implique une estimation des fonctions de transfert acoustiques $\boldsymbol{H_i}$ entre la source de parole et chacun des micros M_i (M_1, M_3 ou M_4).

[0075] Ces fonctions de transfert sont avantageusement évaluées par un estimateur de type LMS fréquentiel (bloc 30) recevant en entrée les signaux issus des différents micros et délivrant en sortie les estimées des diverses fonctions de transfert ***H.***

[0076] Il est également nécessaire d'estimer (bloc 32) la matrice de corrélation $\boldsymbol{R_n}$ (matrice spectrale de covariance, également dénommée matrice interspectrale des bruits).

[0077] Enfin, ces diverses estimations impliquent la connaissance d'une probabilité de présence de parole *SPP*, obtenue à partir du signal recueilli par l'un des micros (bloc 34).

[0078] On va maintenant décrire en détail la manière dont opère l'estimateur MMSE-STSA.

[0079] Il s'agit de traiter les signaux multiples produits par les micros pour fournir un signal débruité unique qui soit le plus proche possible du signal de parole émis par le locuteur, c'est-à-dire :

- contenant le moins de bruit possible, et
- déformant le moins possible la voix du locuteur restituée en sortie.

[0080] Sur le micro de rang *i*, le signal recueilli est :

$$x_i(t) = h_i \otimes s(t) + b_i(t)$$

où x_i est le signal capté, h_i est la réponse impulsionnelle entre la source de signal utile (signal de parole du locuteur) et le micro M_i, *s* est le signal utile produit par la source S et b_i est le bruit additif.

[0081] Pour l'ensemble des micros, on peut utiliser la notation vectorielle :

$$\boldsymbol{x}(t) = \boldsymbol{h} \otimes s(t) + \boldsymbol{b}(t)$$

[0082] Dans le domaine fréquentiel, cette expression devient (les majuscules représentant les transformées de Fourier correspondantes) :

$$X_i(\omega) = H_i(\omega)S(\omega) + B_i(\omega)$$

[0083] On fera les hypothèses suivantes, pour toutes les fréquences ω:

- le signal $S(\omega)$ est gaussien de moyenne nulle et de puissance spectrale $\sigma_s^2(\omega)$;
- les bruits $B_i(\omega)$ sont gaussiens de moyenne nulle et ont une matrice interspectrale ($E[BB^T]$) notée $\Sigma_{bb}(\omega)$;
- le signal et les bruits considérés sont décorrélés, et chacun est décorrélé lorsque les fréquences sont différentes.

[0084] Comme cela a été indiqué plus haut, dans le cas multi-microphone l'estimateur MMSE-STSA se factorise en un *beamforming* MVDR (bloc 28) suivi d'un estimateur monocanal (l'algorithme OM/LSA du bloc 24).

[0085] Le *beamforming* MVDR s'écrit :

$$MVDR(\mathbf{X}) = \frac{\mathbf{H}^T \Sigma_{bb}^{-1} \mathbf{X}}{\mathbf{H}^T \Sigma_{bb}^{-1} \mathbf{H}}$$

[0086] Le beamforming MVDR adaptatif exploite ainsi la cohérence du signal utile pour estimer une fonction de transfert ***H*** correspondant au canal acoustique entre le locuteur et chacun des micros du sous-réseau.

[0087] Pour l'estimation de ce canal acoustique, on utilise un algorithme de type bloc-LMS dans le domaine fréquentiel (bloc 30) tel que celui décrit notamment par :

[4] J. Prado and E. Moulines, *Frequency-Domain Adaptive Filtering with Applications to Acoustic Echo Cancellation,* Springer, Ed. Annals of Telecommunications, 1994.

[0088] Les algorithmes de type LMS - ou NLMS (*Normalized LMS*) qui est une version normalisée du LMS - sont des algorithmes relativement simples et peu exigeants en termes de ressources de calcul.

[0089] Pour un *beamforming* de type GSC (*Generalized Sidelobe Canceller*), cette approche est similaire à celle proposée par :

[5] M.-S. Choi, C.-H. Baik, Y.-C. Park, and H.-G. Kang, "A Soft-Decision Adaptation Mode Controller for an Efficient Frequency-Domain Generalized Sidelobe Canceller," *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP* 2007, Vol. 4, April 2007, pp. IV-893-IV-896.

[0090] Le signal utile $s(t)$ étant inconnu, on ne peut identifier ***H*** qu'à une fonction de transfert près. On choisit donc l'un des canaux comme référence de signal utile, par exemple le canal du micro M_1, et l'on calcule les fonctions de transfert $\boldsymbol{H}_2$ ... $\boldsymbol{H}_n$ pour les autres canaux (ce qui revient à contraindre $\boldsymbol{H}_1 = 1$). Si le micro de référence choisi n'apporte pas de dégradation majeure sur le signal utile, ce choix n'a pas d'influence notable sur les performances de l'algorithme.

[0091] Comme illustré sur la figure, l'algorithme LMS vise (de façon connue) à estimer un filtre ***H*** (bloc 36) au moyen d'un algorithme adaptatif, correspondant au signal x_i délivré par le micro M_1, en estimant le transfert de voix entre le micro M_i et le micro M_1 (pris comme référence). La sortie du filtre 36 est soustraite en 38 au signal x_1 capté par le micro M_1, pour donner un signal d'erreur de prédiction permettant l'adaptation itérative du filtre 36. Il est ainsi possible de prédire à partir du signal x_i la composante de parole contenue dans le signal x_1.

[0092] Pour éviter les problèmes liés à la causalité (c'est-à-dire pour être sûr que les signaux x_i arrivent en avance par rapport à la référence x_1), on retarde légèrement (bloc 40) le signal x_1.

[0093] Par ailleurs, on pondère en 42 le signal d'erreur du filtre adaptatif 36 par la probabilité de présence de parole *SPP* délivrée en sortie du bloc 34, de manière à ne procéder à l'adaptation du filtre que quand la probabilité de présence de parole est élevée.

[0094] Cette pondération peut notamment être opérée par modification du pas d'adaptation de l'algorithme, en fonction

de la probabilité *SPP.*

[0095] L'équation de mise à jour du filtre adaptatif est, pour le *bin* de fréquence *k* et pour le micro *i* :

$$H_i(t,k) = H_i(t-1,k) + \mu X_1(t,k)^*(X_1(t,k) - H_i(t-1,k)\, X_i(t,k))$$

avec

$$\mu = \mu_0 \frac{SPP(t,k)}{E[|X_1(k)|^2]}$$

t étant l'indice temporel de la trame courante, μ_0 étant une constante choisie expérimentalement, et *SPP* étant la probabilité de présence de parole *a posteriori,* estimée comme indiqué plus haut (bloc 34).

[0096] Le pas μ d'adaptation de l'algorithme, modulé par la probabilité de présence de parole *SPP,* s'écrit sous forme normalisée du LMS (le dénominateur correspondant à la puissance spectrale du signal x_1 à la fréquence considérée) :

$$\mu = \frac{p}{E[\boldsymbol{X_1^2}]}$$

[0097] L'hypothèse que les bruits sont décorrélés conduit à une prédiction de la voix, et non du bruit, par l'algorithme LMS, de sorte que la fonction de transfert estimé correspond effectivement au canal acoustique ***H*** entre le locuteur et les micros.

Algorithme SDW-MWF de débruitage BF (bloc 18)

[0098] On va décrire un exemple d'algorithme de débruitage du type SDW-MWF, opéré dans le domaine temporel, mais ce choix n'est pas limitatif, d'autres techniques de débruitage pouvant être envisagées, dès lors qu'elles sont basées sur la prédiction du bruit d'un micro sur l'autre. En outre, ce débruitage BF n'est pas nécessairement opéré dans le domaine temporel, il peut également être opéré dans le domaine fréquentiel, par des moyens équivalents.

[0099] La technique employée par l'invention est basée sur une prédiction du bruit d'un micro sur l'autre décrite, pour une aide auditive, par :

> [6] A. Spriet, M. Moonen, and J. Wouters, "Stochastic Gradient-Based Implementation of Spatially Preprocessed Speech Distortion Weighted Multichannel Wiener Filtering for Noise Reduction in Hearing Aids," IEEE Transactions on Signal Processing, Vol. 53, pp. 911-925, Mar. 2005.

[0100] Chaque micro capte une composante de signal utile et une composante de bruit. Pour le micro de rang *i*, on a : $x_i(t) = s_i(t) + b_i(t)$ s_i étant la composante du signal utile et b_i la composante de bruit. Si l'on souhaite estimer une version du signal utile présente sur un micro *k* par un estimateur des moindres carrés linéaires, ceci revient à estimer un filtre **W** de taille *M.L* tel que :

$$\hat{\mathbf{W}}_k = \min_{\mathbf{w}} E\left[\left|s_k(t) - \mathbf{w}^T \mathbf{x}(t)\right|^2\right]$$

où :

$\mathbf{x}_i(t)$ est le vecteur $[x_i(t - L + 1) \ldots x_i(t)]^T$ et

$$\mathbf{x}(t) = [\mathbf{x}_1(t)^T \quad \mathbf{x}_2(t)^T \quad \dots \quad \mathbf{x}_M(t)^T]^T.$$

[0101] La solution est donnée par le filtre de Wiener :

$$\hat{\mathbf{W}}_k = \left[E\left[\mathbf{x}(t)\mathbf{x}(t)^T\right]\right]^{-1} E\left[\mathbf{x}(t)s_k(t)\right]$$

[0102] Dans la mesure où, comme on l'a expliqué en introduction, pour la partie BF du spectre on cherche à estimer le bruit et non plus le signal utile, on obtient :

$$\hat{\mathbf{W}}_k^b = \min_{\mathbf{w}} E\left[\left|b_k(t) - \mathbf{w}^T\mathbf{x}(t)\right|^2\right]$$

[0103] Cette prédiction du bruit présent sur un micro est opérée à partir du bruit présent sur tous les micros considérés du second sous-réseau R_2, et ceci dans les périodes de silence du locuteur, où seul le bruit est présent.

[0104] La technique utilisée est voisine de celle du débruitage ANC (*Adaptative Noise Cancellation*), en utilisant plusieurs micros pour la prédiction et en incluant dans le filtrage un micro de référence (par exemple le micro M_1). La technique ANC est exposée notamment par :

[7] B. Widrow, J. Glover, J.R., J. McCool, J. Kaunitz, C. Williams, R. Hearn, J. Zeidler, J. Eugene Dong, and R. Goodlin, "Adaptive Noise Cancelling : Principles and applications," Proceedings of the IEEE, Vol. 63, No. 12, pp. 1692-1716, Dec. 1975.

[0105] Comme illustré sur la Figure 3, le filtre de Wiener (bloc 44) fournit une prédiction du bruit qui est soustraite en 46 du signal recueilli, non débruité, après application d'un retard (bloc 48) pour éviter les problèmes de causalité. Le filtre de Wiener 44 est paramétré par un coefficient μ (schématisé en 50) qui détermine une pondération ajustable entre, d'une part, la distorsion introduite par le traitement sur le signal vocal débruité et, d'autre part, le niveau de bruit résiduel.

[0106] Dans le cas d'un signal recueilli par un plus grand nombre de micros, la généralisation de ce schéma de la prédiction de bruit pondéré est donnée Figure 4.

[0107] Le signal estimé étant :

$$\hat{s}(t) = x_k(t) - \mathbf{W}_k^b$$

la solution est donnée, de la même façon que précédemment, par le filtre de Wiener:

$$\hat{\mathbf{W}}_k^b = \left[E\left[\mathbf{x}(t)\mathbf{x}(t)^T\right]\right]^{-1} E\left[\mathbf{x}(t)b_k(t)\right]$$

[0108] Le signal estimé est alors rigoureusement le même, car on peut prouver que $\hat{\mathbf{W}}_k + \hat{\mathbf{W}}_k^b = \mathbf{e}_k$, avec $\mathbf{e}_k = [0 \quad 0 \quad \dots \underbrace{1}_{\text{position k}} \dots \quad 0]^T$.

[0109] Le filtre de Wiener utilisé est avantageusement un filtre de Wiener pondéré (SDW-MVF), pour prendre en compte non seulement l'énergie du bruit à éliminer par le filtrage, mais également la distorsion introduite par ce filtrage et qu'il convient de minimiser.

[0110] Dans le cas du filtre de Wiener $\hat{W}_k$, la "fonction de cout" peut être séparée en deux, l'écart quadratique moyen

pouvant s'écrire comme la somme de deux termes :

$$E\left[\left|s_k(t) - \mathbf{w}^T \mathbf{x}(t)\right|^2\right] = \underbrace{E\left[\left|s_k(t) - \mathbf{w}^T \mathbf{s}(t)\right|^2\right]}_{e_s} + \underbrace{E\left[\left|\mathbf{w}^T \mathbf{b}(t)\right|^2\right]}_{e_b}$$

où :

$s_i(t)$ est le vecteur $[s_i(t - L + 1) \ldots s_i(t)]^T$

- $\mathbf{s}(t) = [\mathbf{s}_1(t)^T \mathbf{s}_2(t)^T \ldots \mathbf{s}_M(t)^T]^T$
- $\mathbf{b}_i(t)$ est le vecteur $[b_i(t - L + 1) \ldots b_i(t)]^T$ et
- $\mathbf{b}(t) = [\mathbf{b}_1(t)^T \mathbf{b}_2(t)^T \ldots \mathbf{b}_M(t)^T]^T$
- e_s est la distorsion introduite par le filtrage sur le signal utile, et
- e_b est le bruit résiduel après filtrage.

[0111] Il est possible de pondérer ces deux erreurs e_s et e_b selon que l'on privilégie la réduction de distorsion ou bien la réduction du bruit résiduel.

[0112] En invoquant la décorrélation entre bruit et signal utile, le problème devient :

$$\hat{\mathbf{W}}_{kr} = \min_{\mathbf{w}} \left[E\left[\left|s_k(t) - \mathbf{w}^T \mathbf{s}(t)\right|^2\right]\right] + \left[\mu E\left[\left|\mathbf{w}^T \mathbf{b}(t)\right|^2\right]\right]$$

avec pour solution :

$$\hat{\mathbf{W}}_{kr} = \left[E\left[\mathbf{s}(t)\mathbf{s}(t)^T\right] + \mu E\left[\mathbf{b}(t)\mathbf{b}(t)^T\right]\right]^{-1} E\left[\mathbf{s}(t)s_k(t)\right]$$

l'indice "$_r$' indiquant que l'on régularise la fonction de cout pour pondérer selon la distorsion, et μ étant un paramètre ajustable :

- plus μ est grand, plus l'on privilégie la réduction du bruit, mais au prix d'une distorsion plus importante sur le signal utile ;
- si μ est nul, aucune importance n'est accordée à la réduction du bruit, et la sortie vaut $x_k(t)$ car les coefficients du filtre sont nuls ;
- si μ est infini, les coefficients du filtre sont nuls à l'exception du terme en position $k*L$ (L étant la longueur du filtre) qui vaut 1, la sortie vaut donc zéro.

[0113] Pour le filtre dual $\mathbf{W}_k^b$, le problème peut se réécrire :

$$\hat{\mathbf{W}}_{kr}^b = \min_{\mathbf{w}} \mu\left[E\left[\left|b_k(t) - \mathbf{w}^T \mathbf{b}(t)\right|^2\right]\right] + \left[E\left[\left|\mathbf{w}^T \mathbf{s}(t)\right|^2\right]\right]$$

avec pour solution :

$$\hat{\mathbf{W}}_{kr}^{b} = \left[\frac{1}{\mu}E\left[\mathbf{s}(t)\mathbf{s}(t)^{T}\right] + E\left[\mathbf{b}(t)\mathbf{b}(t)^{T}\right]\right]^{-1} E\left[\mathbf{b}(t)b_{k}(t)\right]$$

[0114] On démontre également que le signal de sortie est le même quelle que soit l'approche utilisée.

[0115] Ce filtre est mis en oeuvre de manière adaptative, par un algorithme de descente de gradient tel que celui exposé dans l'article [6] précité.

[0116] Le schéma est celui illustré Figures 3 et 4.

[0117] Pour la mise en oeuvre de ce filtre, il est nécessaire d'estimer les matrices $R_s = E[s(t)s(t)^T]$, $R_b = E[b(t)b(t)^T]$, le vecteur $E[b(t)b_k(t)]$ ainsi que les paramètres L (la longueur souhaitée pour le filtre) et μ (qui ajuste la pondération entre réduction de bruit et distorsion).

[0118] Si l'on suppose que l'on dispose d'un détecteur d'activité vocale (qui permet de discriminer entre phases de parole du locuteur et phases de silence) et que le bruit $b(t)$ est stationnaire, on peut estimer $\mathbf{R}_b$ durant les phases de silence, où seul le bruit est capté par les micros. Pendant ces phases de silence, on estime la matrice $\mathbf{R}_b$ au fil de l'eau :

$$\mathrm{R}_b(t) = \begin{cases} \lambda \mathrm{R}_b(t-1) + (1-\lambda)\mathbf{x}(t)\mathbf{x}(t)^T & \text{s'il n'y a pas de parole} \\ \mathrm{R}_b(t-1) & \text{sinon} \end{cases}$$

λ étant un facteur d'oubli.

[0119] On peut estimer $E[b(t)b_k(t)]$, ou remarquer que c'est une colonne de $\mathbf{R}_b$. Pour estimer $\mathbf{R}_s$, on invoque la décorrélation du bruit et du signal utile. Si l'on note $R_x = E[x(t)x(t)^T]$, on peut alors écrire : R_x, $R_s + R_b$.

[0120] On peut estimer $\mathbf{R}_x$ de la même façon que $\mathbf{R}_b$, mais sans condition sur la présence de parole :

$$\mathrm{R}_x(t) = \lambda \mathrm{R}_x(t-1) + (1-\lambda)\mathbf{x}(t)\mathbf{x}(t)^T$$

ce qui permet de déduire $R_s(t) = R_x(t) - R_b(t)$.

[0121] En ce qui concerne la longueur L du filtre, ce paramètre doit correspondre à une réalité spatiale et temporelle, avec un nombre de coefficients suffisant pour prédire le bruit temporellement (cohérence temporelle du bruit) et spatialement (transfert spatial entre les micros).

[0122] Le paramètre μ est ajusté expérimentalement, en l'augmentant jusqu'à ce que la distorsion sur la voix devienne perceptible à l'oreille.

[0123] Ces estimateurs sont utilisés pour opérer une descente de gradient sur la fonction de cout suivante :

$$J_{kr} = \mu\left[E\left[\left|b_k(t) - \mathbf{w}^T\mathbf{b}(t)\right|^2\right]\right] + \left[E\left[\left|\mathbf{w}^T\mathbf{s}(t)\right|^2\right]\right]$$

[0124] Le gradient de cette fonction vaut :

$$\delta J_{kr} = 2\left[\mathrm{R}_s + \mu\mathrm{R}_b\right]\mathbf{w} - 2\mu E\left[\mathbf{b}(t)b_k(t)\right]$$

[0125] D'où l'équation de mise à jour :

$$\mathbf{w}(t) = \mathbf{w}(t-1) - \alpha\delta J_{kr}$$

où α est un pas d'adaptation proportionnel à $\frac{1}{\mathbf{x}^T\mathbf{x}}$.

Revendications

1. Un procédé de débruitage d'un signal acoustique bruité pour un dispositif audio multi-microphone opérant dans un milieu bruité,
le signal acoustique bruité comprenant une composante utile issue d'une source de parole et une composante parasite de bruit,
ledit dispositif comprenant un réseau de capteurs formé d'une pluralité de capteurs microphoniques (M_1 ... M_4) disposés selon une configuration prédéterminée et aptes à recueillir le signal bruité,
les capteurs étant regroupés en deux sous-réseaux, avec un premier sous-réseau (R_1) de capteurs apte à recueillir une partie haute fréquence du spectre, et un second sous-réseau (R_2) de capteurs apte à recueillir une partie basse fréquence du spectre distincte de ladite partie haute fréquence,
ce procédé comportant les étapes suivantes :

a) partition du spectre du signal bruité en ladite partie haute fréquence (HF) et ladite partie basse fréquence (BF), par filtrage (10, 16) respectivement au-delà et en deçà d'une fréquence pivot prédéterminée,
b) débruitage de chacune des deux parties du spectre avec mise en oeuvre d'un estimateur à algorithme adaptatif ; et
c) reconstruction du spectre par combinaison (22) des signaux délivrés après débruitage des deux parties du spectre aux étapes b1) et b2),

procédé **caractérisé en ce que** l'étape b) de débruitage est opérée par des traitements distincts pour chacune des deux parties du spectre, avec :

b1) pour la partie haute fréquence, un débruitage exploitant le caractère prédictible du signal utile d'un capteur sur l'autre entre capteurs du premier sous-réseau, au moyen d'un premier estimateur (14) à algorithme adaptatif, et
b2) pour la partie basse fréquence, un débruitage par prédiction du bruit d'un capteur sur l'autre entre capteurs du second sous-réseau, au moyen d'un second estimateur (18) à algorithme adaptatif.

2. Le procédé de la revendication 1, dans lequel le premier sous-réseau de capteurs (R_1) apte à recueillir la partie haute fréquence du spectre comprend un réseau linaire d'au moins deux capteurs (M_1, M_3, M_4) alignés perpendiculairement à la direction (Δ) de la source de parole.

3. Le procédé de la revendication 1, dans lequel le second sous-réseau de capteurs (R_2) apte à recueillir la partie basse fréquence du spectre comprend un réseau linaire d'au moins deux capteurs (M_1, M_2) alignés parallèlement à la direction (Δ) de la source de parole.

4. Le procédé de la revendication 2, dans lequel les capteurs (M_1, M_3, M_4) du premier sous-réseau de capteurs (R_1) sont des capteurs unidirectionnels orientés dans la direction (Δ) de la source de parole.

5. Le procédé de la revendication 2, dans lequel le traitement de débruitage de la partie haute fréquence du spectre à l'étape b1) est opéré de façon différenciée pour une bande inférieure et une bande supérieure de cette partie haute fréquence, avec sélection de capteurs différents parmi les capteurs du premier sous-réseau (R_1), la distance entre les capteurs (M_1, M_4) sélectionnés pour le débruitage de la bande supérieure étant plus réduite que celle des capteurs (M_3, M_4) sélectionnés pour le débruitage de la bande inférieure.

6. Le procédé de la revendication 1 comprenant en outre, après l'étape c) de reconstruction du spectre, une étape de :

d) réduction sélective du bruit (24) par un traitement de type gain à amplitude log-spectrale modifié optimisé, OM-LSA, à partir du signal reconstruit produit à l'étape c) et d'une probabilité de présence de parole.

7. Le procédé de la revendication 1 dans lequel l'étape b1) de débruitage de la partie haute fréquence, exploitant le caractère prédictible du signal utile d'un capteur sur l'autre, est opérée dans le domaine fréquentiel.

8. Le procédé de la revendication 7 dans lequel l'étape b1) de débruitage de la partie haute fréquence, exploitant le caractère prédictible du signal utile d'un capteur sur l'autre, est opérée par :

b11) estimation (34) d'une probabilité de présence de parole (SPP) dans le signal bruité recueilli ;
b12) estimation (32) d'une matrice spectrale de covariance des bruits recueillis par les capteurs du premier sous-réseau, cette estimation étant modulée par la probabilité de présence de parole ;
b13) estimation (30) de la fonction de transfert des canaux acoustiques entre la source de parole et au moins certains des capteurs du premier sous-réseau, cette estimation étant opérée par rapport à une référence de signal utile constituée par le signal recueilli par l'un des capteurs du premier sous-réseau, et étant en outre modulée par la probabilité de présence de parole ; et
b14) calcul (28) d'un projecteur linéaire optimal donnant un signal combiné débruité unique à partir des signaux recueillis par au moins certains des capteurs du premier sous-réseau, de la matrice spectrale de covariance estimée à l'étape b12), et des fonctions de transfert estimées à l'étape b13).

9. Le procédé de la revendication 8, dans lequel l'étape b14) de calcul d'un projecteur linéaire optimal (28) est mise en oeuvre par un estimateur de type *beamforming* à réponse sans distorsion à variance minimale, MVDR.

10. Le procédé de la revendication 9, dans lequel l'étape b13) d'estimation de la fonction de transfert des canaux acoustiques (30) est mise en oeuvre par un filtre adaptatif (36, 38, 40) à prédiction linéaire de type moindres carrés moyens, LMS, avec modulation (42) par la probabilité de présence de parole.

11. Le procédé de la revendication 10, dans lequel ladite modulation par la probabilité de présence de parole est une modulation par variation du pas d'itération du filtre adaptatif LMS.

12. Le procédé de la revendication 1 dans lequel, pour le débruitage de la partie basse fréquence à l'étape b2), la prédiction du bruit d'un capteur sur l'autre est opérée dans le domaine temporel.

13. Le procédé de la revendication 12, dans lequel la prédiction du bruit d'un capteur sur l'autre est mise en oeuvre par un filtre (44, 46, 48) de type filtre de Wiener multicanal avec pondération par la distorsion de la parole, SDW-MWF.

14. Le procédé de la revendication 13, dans lequel le filtre SDW-MWF est estimé de manière adaptative par un algorithme de descente de gradient.

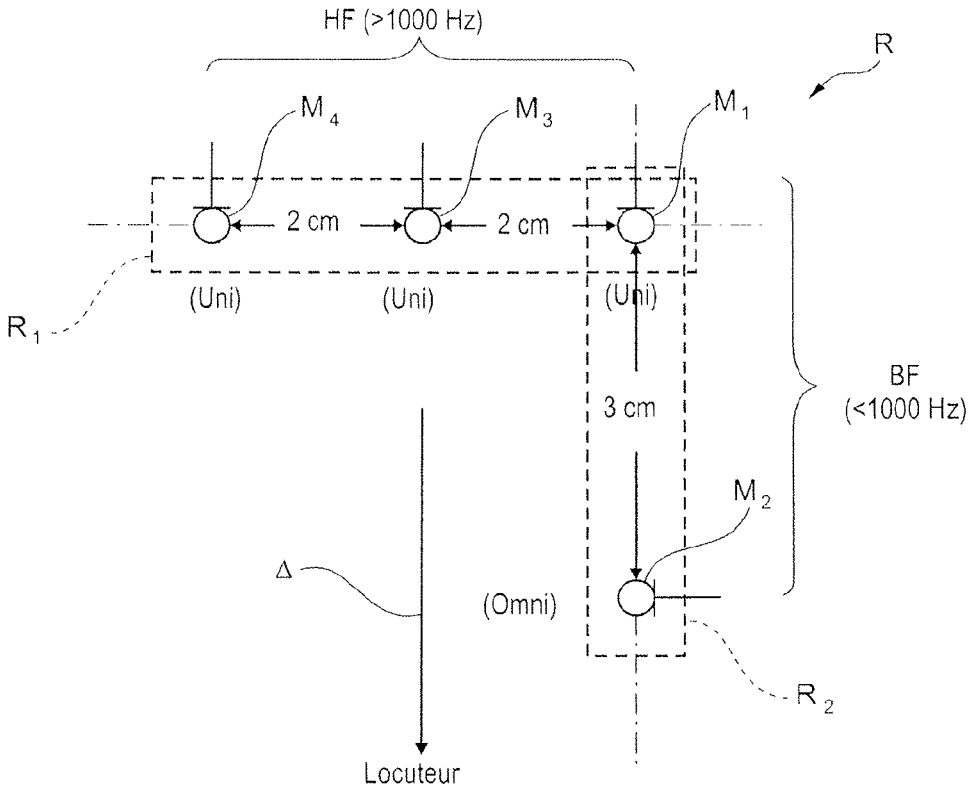
HF (>1000 Hz)
R
M4
M3
M1
2 cm
2 cm
(Uni)
(Uni)
(Uni)
R1
3 cm
BF
(<1000 Hz)
M2
Δ
(Omni)
R2
Locuteur

Fig. 1

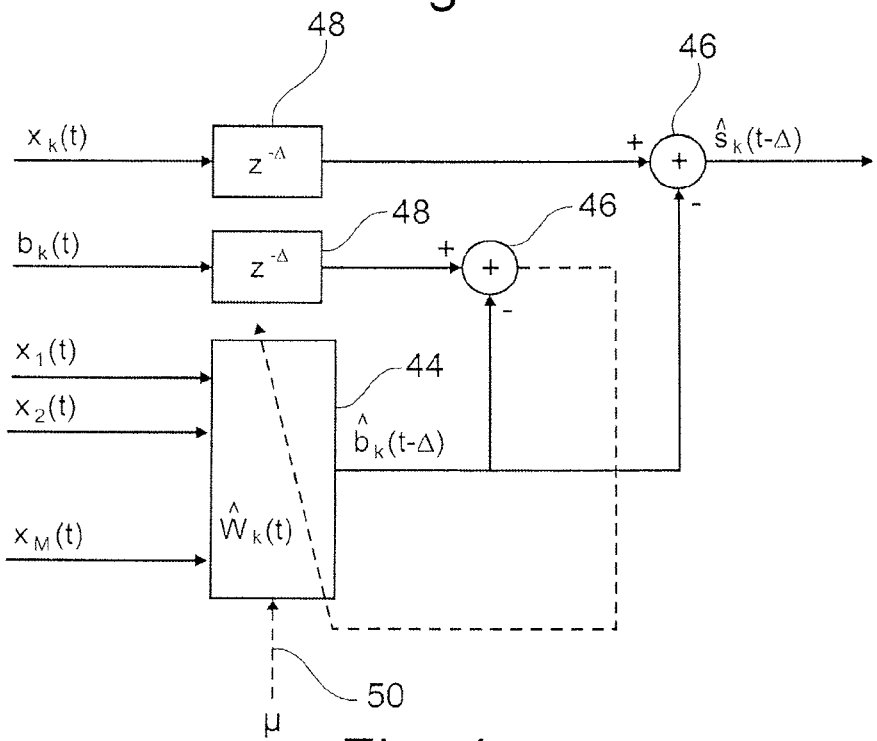
48
46
xk(t)
z-Δ
+
+
ŝk(t-Δ)
-
48
46
bk(t)
z-Δ
+
+
-
x1(t)
x2(t)
44
b̂k(t-Δ)
Ŵk(t)
xM(t)
50
μ

Fig. 4

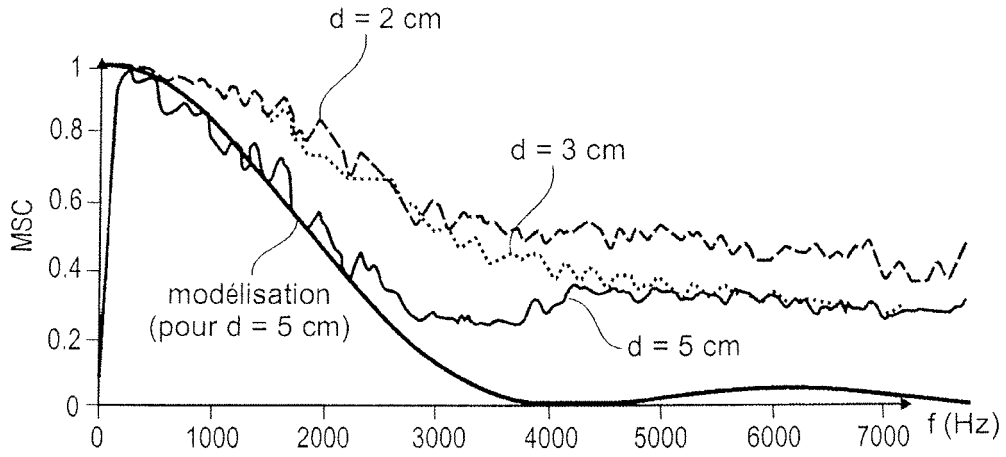

Fig. 2a
(omnidirectionnel)

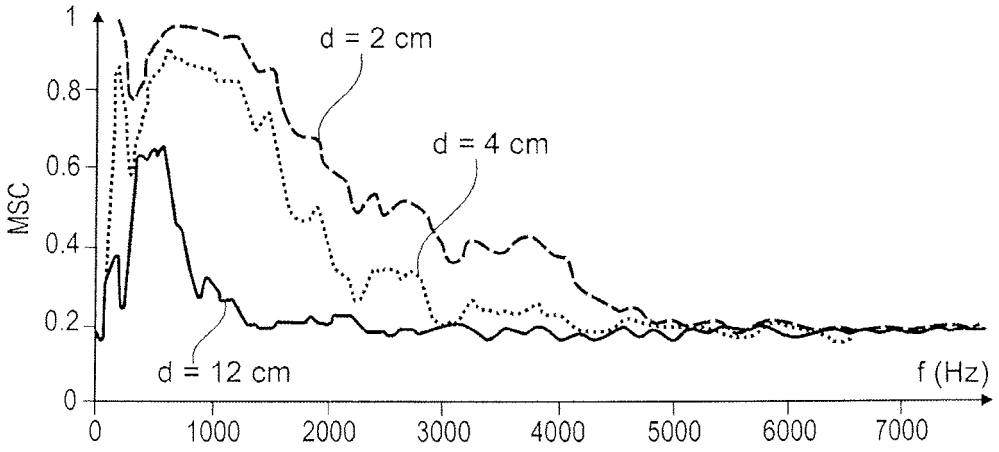

Fig. 2b
(unidirectionnel)

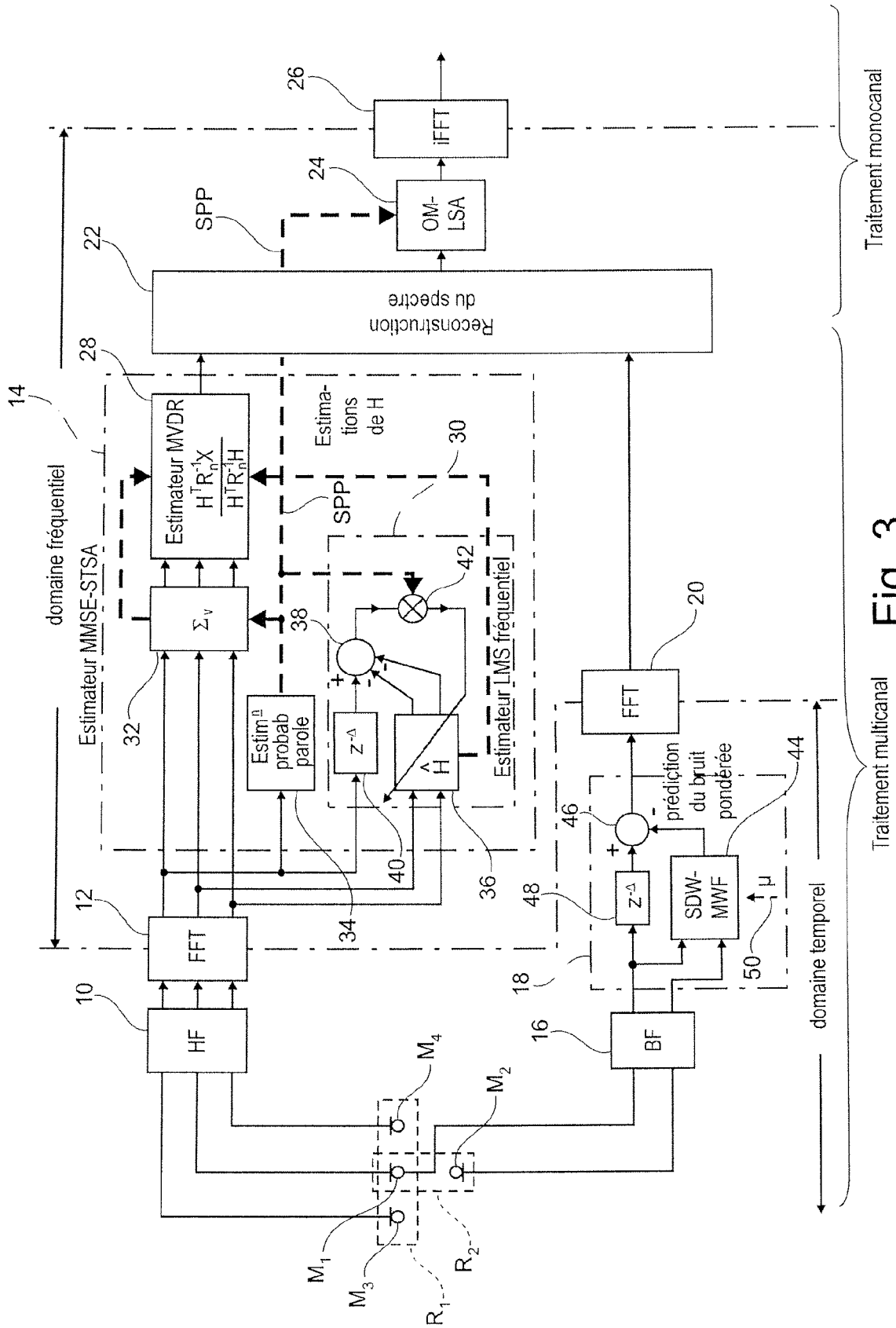
Traitement monocanal
Traitement multicanal
domaine fréquentiel
domaine temporel
Estimateur MMSE-STSA
Estimateur LMS fréquentiel
Estima-tions de H
Reconstruction du spectre
Estimateur MVDR
$\frac{H^T R_n^{-1} X}{H^T R_n^{-1} H}$
Σ_v
Estim^n probab parole
SPP
OM-LSA
iFFT
FFT
HF
BF
SDW-MWF
prédiction du bruit pondérée
$z^{-\Delta}$
$\hat{H}$
μ
M_1
M_2
M_3
M_4
R_1
R_2
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
46
48
50

Fig. 3

Europäisches Patentamt
European Patent Office
Office européen des brevets

RAPPORT DE RECHERCHE EUROPEENNE

Numéro de la demande
EP 13 17 1948

DOCUMENTS CONSIDERES COMME PERTINENTS

Catégorie	Citation du document avec indication, en cas de besoin, des parties pertinentes	Revendication concernée	CLASSEMENT DE LA DEMANDE (IPC)
A	MCCOWAN I A ET AL: "Adaptive parameter compensation for robust hands-free speech recognition using a dual beamforming microphone array", 2 mai 2001 (2001-05-02), INTELLIGENT MULTIMEDIA, VIDEO AND SPEECH PROCESSING, 2001. PROCEEDINGS OF 2001 INTERNATIONAL SYMPOSIUM ON 2-4 MAY 2001, PISCATAWAY, NJ, USA,IEEE, PAGE(S) 547 - 550, XP010544783, ISBN: 978-962-85766-2-3 * page 548, alinéa 2.1 - alinéa 2.2; figures 1,2 *	1-14	INV. G10L21/0208 G10L21/0216
A	US 2003/040908 A1 (YANG FENG [US] ET AL) 27 février 2003 (2003-02-27) * page 4, alinéa 52 - alinéa 53; revendication 1; figures 6, 8 * * page 7, alinéa 88 - alinéa 92 *	1-14	
A	WO 2008/104446 A2 (PHONAK AG [CH]; KORL SASCHA [CH]; DERLETH RALPH PETER [CH]; SCHUSTER G) 4 septembre 2008 (2008-09-04) * page 11, ligne 7 - ligne 23; revendications 1,3; figures 3,6 *	1-14	DOMAINES TECHNIQUES RECHERCHES (IPC) G10L
A	EP 1 640 971 A1 (HARMAN BECKER AUTOMOTIVE SYS [DE]) 29 mars 2006 (2006-03-29) * alinéa [0026] - alinéa [0033]; revendication 1; figure 2 *	1-14	

Le présent rapport a été établi pour toutes les revendications

Lieu de la recherche	Date d'achèvement de la recherche	Examinateur
La Haye	18 septembre 2013	Perez Molina, E

CATEGORIE DES DOCUMENTS CITES

X : particulièrement pertinent à lui seul
Y : particulièrement pertinent en combinaison avec un autre document de la même catégorie
A : arrière-plan technologique
O : divulgation non-écrite
P : document intercalaire

T : théorie ou principe à la base de l'invention
E : document de brevet antérieur, mais publié à la date de dépôt ou après cette date
D : cité dans la demande
L : cité pour d'autres raisons
.....
& : membre de la même famille, document correspondant

4

EPO FORM 1503 03.82 (P04C02)

ANNEXE AU RAPPORT DE RECHERCHE EUROPEENNE RELATIF A LA DEMANDE DE BREVET EUROPEEN NO. EP 13 17 1948

La présente annexe indique les membres de la famille de brevets relatifs aux documents brevets cités dans le rapport de recherche européenne visé ci-dessus.
Lesdits members sont contenus au fichier informatique de l'Office européen des brevets à la date du
Les renseignements fournis sont donnés à titre indicatif et n'engagent pas la responsabilité de l'Office européen des brevets.

18-09-2013

Document brevet cité au rapport de recherche	Date de publication	Membre(s) de la famille de brevet(s)	Date de publication
US 2003040908 A1	27-02-2003	AUCUN	
WO 2008104446 A2	04-09-2008	AT 551692 T	15-04-2012
		EP 2238592 A2	13-10-2010
		US 2010329492 A1	30-12-2010
		WO 2008104446 A2	04-09-2008
EP 1640971 A1	29-03-2006	AT 405925 T	15-09-2008
		CA 2518684 A1	23-03-2006
		CN 1753084 A	29-03-2006
		EP 1640971 A1	29-03-2006
		JP 4734070 B2	27-07-2011
		JP 2006094522 A	06-04-2006
		KR 20060051582 A	19-05-2006
		US 2006222184 A1	05-10-2006

EPO FORM P0460

Pour tout renseignement concernant cette annexe : voir Journal Officiel de l'Office européen des brevets, No.12/82

RÉFÉRENCES CITÉES DANS LA DESCRIPTION

Cette liste de références citées par le demandeur vise uniquement à aider le lecteur et ne fait pas partie du document de brevet européen. Même si le plus grand soin a été accordé à sa conception, des erreurs ou des omissions ne peuvent être exclues et l'OEB décline toute responsabilité à cet égard.

Documents brevets cités dans la description

- US 20080280653 A1 **[0010]**
- US 20070165879 A1 **[0012]**
- EP 2293594 A1 **[0014]**
- EP 2309499 A1, Parrot **[0014]**
- WO 2007099222 A1, Parrot **[0062]**

Littérature non-brevet citée dans la description

- **I. MCCOWAN ; S. SRIDHARAN.** Adaptive Parameter Compensation for Robust Hands-free Speech Recognition using a Dual-Beamforming Microphone Array. *Proceedings on 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing,* Mai 2001 **[0015]**
- **I. COHEN.** Optimal Speech Enhancement under Signal Presence Uncertainty Using Log-Spectral Amplitude Estimator. *Signal Processing Letters,* Avril 2002, vol. 9 (4), 113-116 **[0060]**
- **I. COHEN ; B. BERDUGO.** Two-Channel Signal Detection and Speech Enhancement Based on the Transient Beam-to-Reference Ratio. *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2003,* Avril 2003, 233-236 **[0061]**
- **R. C. HENDRIKS et al.** On optimal multichannel mean-squared error estimators for speech enhancement. *IEEE Signal Processing Letters,* 2009, vol. 16 (10 **[0071]**
- **A. SPRIET ; M. MOONEN ; J. WOUTERS.** Stochastic Gradient-Based Implementation of Spatially Preprocessed Speech Distortion Weighted Multichannel Wiener Filtering for Noise Reduction in Hearing Aids. *IEEE Transactions on Signal Processing,* Mars 2005, vol. 53, 911-925 **[0099]**
- **B. WIDROW ; J. GLOVER, J.R. ; J. MCCOOL ; J. KAUNITZ ; C. WILLIAMS ; R. HEARN ; J. ZEIDLER ; J. EUGENE DONG ; R. GOODLIN.** Adaptive Noise Cancelling : Principles and applications. *Proceedings of the IEEE,* Décembre 1975, vol. 63 (12), 1692-1716 **[0104]**