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

(11) **EP 2 530 673 A1**

(12) **DEMANDE DE BREVET EUROPEEN**

(43) Date de publication:
05.12.2012 Bulletin 2012/49

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

(21) Numéro de dépôt: **12170407.6**

(22) Date de dépôt: **01.06.2012**

(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: **01.06.2011 FR 1154825**

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

(72) Inventeurs:
- **Vitte, Guillaume**
  75003 Paris (FR)
- **Herve, Michael**
  75011 Paris (FR)

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

(54) **Equipement audio comprenant des moyens de débruitage d'un signal de parole par filtrage à délai fractionnaire**

(57) L'équipement comprend deux micros (10, 12), des moyens d'échantillonnage et des moyens de débruitage. Les moyens de débruitage sont des moyens de réduction de bruit non fréquentielle comprenant un combineur (14) à filtre adaptatif (16) opérant par recherche itérative visant à annuler le bruit capté par l'un des micros (10) sur la base d'une référence de bruit donnée par l'autre micro (12). Le filtre adaptatif est un filtre à délai fractionnaire modélisant un retard inférieur à la période d'échantillonnage. L'équipement comprend en outre des moyens de détection d'activité vocale (20) délivrant un signal représentatif de la présence ou de l'absence de parole par l'utilisateur de l'équipement. Le filtre adaptatif reçoit en entrée ce signal de manière à, sélectivement : i) soit opérer une recherche adaptative des paramètres du filtre en l'absence de parole, ii) soit figer ces paramètres du filtre en présence de parole.

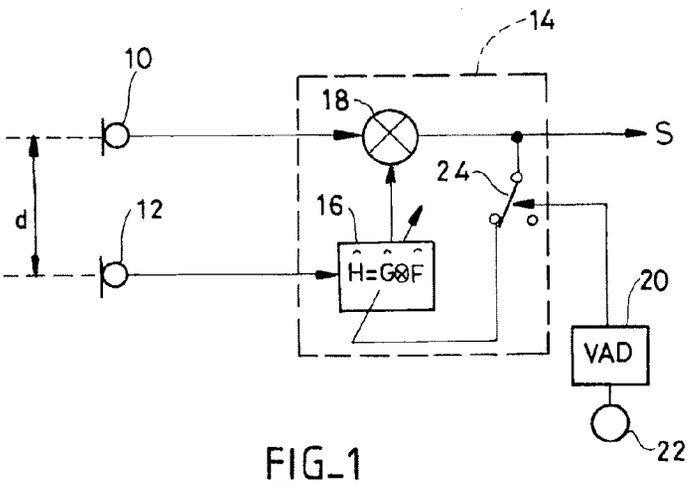

FIG_1

EP 2 530 673 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 boîtier amovible intégrant tous les composants et fonctions de traitement du signal pour la communication téléphonique.

[0005] En effet, la distance importante entre le micro (placé au niveau de la planche de bord ou dans un angle supérieur du pavillon de l'habitacle) et le locuteur (dont l'éloignement est contraint par la position de conduite) entraîne 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 non stationnaires, c'est-à-dire 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 du même genre se présentent dans le cas où 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 certes 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 plusieurs micros, généralement deux micros, pour combiner de façon judicieuse les signaux captés simultanément par ces micros afin d'isoler les composantes de parole utiles des composantes de bruits parasites.

[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 telle méthode 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] Le problème général de l'invention est, dans un tel contexte, de procéder à une 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.

[0015] Le problème de l'invention est également, dans une telle situation, de pouvoir mettre en oeuvre un ensemble de micros à la fois en nombre réduit (avantageusement deux micros seulement) et relativement rapprochés (typiquement un écartement de quelques centimètres seulement). Un autre aspect important du problème est la nécessité de restituer un signal de parole naturelle et intelligible, c'est-à-dire non distordu et dont le spectre des fréquences utiles ne soit pas amputé par les traitements de débruitage.

[0016] A cet effet, l'invention propose un équipement audio du type général divulgué par le US 2008/0280653 A1 précité, c'est-à-dire comprenant : un ensemble de deux capteurs microphoniques aptes à recueillir la parole de l'utilisateur de l'équipement et à délivrer des signaux de parole bruités respectifs ; des moyens d'échantillonnage des signaux de parole délivrés par les capteurs microphoniques ; et des moyens de débruitage d'un signal de parole, recevant en entrée les échantillons des signaux de parole délivrés par les deux capteurs microphoniques, et délivrant en sortie un signal de parole débruité représentatif de la parole émise par l'utilisateur de l'équipement. Les moyens de débruitage sont des moyens de réduction de bruit non fréquentielle comprenant un combineur à filtre adaptatif des signaux délivrés par les deux capteurs microphoniques, opérant par recherche itérative visant à annuler le bruit capté par l'un des capteurs microphoniques sur la base d'une référence de bruit donnée par le signal délivré par l'autre capteur microphonique.

[0017] De façon caractéristique de l'invention, le filtre adaptatif est un filtre à délai fractionnaire, apte à modéliser un retard inférieur à la période d'échantillonnage des moyens d'échantillonnage. L'équipement comprend en outre des moyens de détection d'activité vocale aptes à délivrer un signal représentatif de la présence ou de l'absence de parole par l'utilisateur de l'équipement, et le filtre adaptatif reçoit également en entrée le signal de présence ou d'absence de parole, de manière à, sélectivement : i) soit opérer une recherche adaptative des paramètres du filtre en l'absence de parole, ii) soit figer ces paramètres du filtre en présence de parole.

[0018] Le filtre adaptatif est notamment apte à estimer un filtre optimal *H* tel que :

$$\hat{H} = \hat{G} \otimes \hat{F}$$

avec :

$$x'(n) = G \otimes x(n) \qquad \text{et} \qquad G(k) = \mathrm{sinc}(k + \tau / Te),$$

- $\hat{H}$ représentant l'estimée du filtre optimal *H*, transfert de bruit entre les deux capteurs microphoniques pour une réponse impulsionnelle incluant un délai fractionnaire,
- $\hat{G}$ représentant l'estimée du filtre à délai fractionnaire *G* entre les deux capteurs microphoniques,
- $\hat{F}$ représentant l'estimée de la réponse acoustique de l'environnement,
- $\otimes$ indiquant une convolution,
- *x(n)* étant la série d'échantillons du signal en entrée du filtre *H*,
- *x'(n)* étant la série *x*(*n*) décalée d'un retard τ,
- *Te* étant la période d'échantillonnage du signal en entrée du filtre *H*,
- τ étant ledit délai fractionnaire, égal à un sous-multiple de *Te,* et
- sinc indiquant la fonction sinus cardinal.

[0019] De préférence, le filtre adaptatif est un filtre à algorithme de prédiction linéaire de type moindres carrés moyens LMS.

[0020] Dans une forme de réalisation, l'équipement comprend une caméra video dirigée vers l'utilisateur de l'équipement et apte à capter une image de celui-ci, et les moyens de détection d'activité vocale comprennent des moyens d'analyse video aptes à analyser l'image produite par la caméra et à délivrer en réponse ledit signal de présence ou d'absence de parole par ledit utilisateur.

[0021] Dans une autre forme de réalisation, l'équipement comprend un capteur physiologique apte à venir en contact avec la tête de l'utilisateur de l'équipement pour y être couplé afin de capter les vibrations vocales non acoustiques transmises par conduction osseuse interne, et les moyens de détection d'activité vocale comprennent des moyens aptes à analyser le signal délivré par le capteur physiologique et à délivrer en réponse ledit signal de présence ou d'absence de parole par ledit utilisateur, notamment par évaluation de l'énergie du signal délivré par le capteur physiologique et comparaison à un seuil.

[0022] L'équipement peut en particulier être un casque audio du type combiné micro/casque, comprenant : des écouteurs comportant chacun un transducteur de reproduction sonore d'un signal audio logé dans une coque pourvue d'un coussinet circumaural ; lesdits deux capteurs microphoniques, disposés sur la coque de l'un des écouteurs ; et ledit

capteur physiologique, incorporé au coussinet de l'un des écouteurs et placé dans une région de celui-ci apte à venir en contact avec la joue ou la tempe du porteur du casque. Ces deux capteurs microphoniques sont de préférence alignés en un réseau linéaire suivant une direction principale dirigée vers la bouche de l'utilisateur de l'équipement.

[0023] 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, sous forme de blocs fonctionnels, la manière dont est réalisé le traitement de débruitage selon l'invention.

La Figure 2 est une représentation graphique de la fonction sinus cardinal modélisée dans le traitement de débruitage de l'invention.

Les Figures 3a et 3b sont des représentations de la fonction sinus cardinal de la Figure 2, respectivement pour les différents points d'une série d'échantillons de signal, et pour la même série décalée dans le temps d'une valeur fractionnaire.

La Figure 4 est une représentation de la réponse acoustique de l'environnement, avec en ordonnée l'amplitude et en abscisse les coefficients du filtre représentant ce transfert.

La Figure 5 est homologue de la Figure 4, après convolution avec une réponse en sinus cardinal.

La Figure 6 est une représentation schématique d'un mode de réalisation consistant à utiliser une caméra pour assurer la détection d'activité vocale.

La Figure 7 illustre de façon générale un ensemble micro/casque combiné auquel peuvent être appliqués les enseignements de l'invention.

La Figure 8 est un schéma d'ensemble qui illustre sous forme de blocs fonctionnels la manière dont peut être réalisé le traitement du signal pour délivrer en sortie un signal débruité représentatif de la parole émise par le porteur du casque de la Figure 7.

La Figure 9 illustre deux chronogrammes correspondant respectivement à un exemple de signal brut recueilli par les micros, et de signal recueilli par un capteur physiologique permettant de distinguer les périodes de parole et les périodes de silence du locuteur.

[0024] La Figure 1 illustre de façon schématique, sous forme de blocs, les différentes fonctions mises en oeuvre par l'invention.

[0025] Le processus de l'invention est mis en oeuvre par des moyens logiciels, schématisés par un certain nombre de blocs fonctionnels correspondant à des algorithmes appropriés exécutés par un microcontrôleur ou un processeur numérique de signal. Bien que, pour la clarté de l'exposé, les différentes fonctions soient présentées sous forme de modules distincts, elles mettent en oeuvre des éléments communs et correspondent en pratique à une pluralité de fonctions globalement exécutées par un même logiciel.

[0026] Le signal que l'on souhaite débruiter est issu d'un réseau de capteurs microphoniques qui, dans la configuration minimale illustrée, peut être simplement un réseau de deux capteurs disposés selon une configuration prédéterminée, chaque capteur étant constitué d'un micro respectif correspondant 10, 12.

[0027] L'invention peut toutefois être généralisée à un réseau de plus de deux capteurs microphoniques, et/ou à des capteurs microphoniques dont chaque capteur est constitué d'une structure plus complexe qu'un simple micro, par exemple une combinaison de plusieurs micros et/ou autres capteurs de parole.

[0028] Les micros 10, 12 sont des micros qui captent le signal émis par la source de signal utile (le signal de parole du locuteur), et la différence de position entre les deux micros induit un ensemble de déphasages et variations d'amplitude dans l'enregistrement des signaux émis par la source de signal utile.

[0029] En pratique, les deux micros 10, 12, sont des micros omnidirectionnels disposés à quelques centimètres l'un de l'autre sur le plafonnier d'un habitacle de voiture, sur la façade d'un autoradio ou d'un emplacement approprié de la planche de bord, ou bien sur la coque d'un des écouteurs d'un casque audio, etc.

[0030] Comme on le verra, la technique de l'invention permet d'assurer un débruitage efficace même pour des micros très rapprochés, c'est-à-dire espacés entre eux d'un écartement *d* tel 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. Ceci correspond à une distance maximale *d* de l'ordre de 4,7 cm pour une fréquence d'échantillonnage F_e de 8 kHz (et un écartement *d* moitié moindre pour une fréquence double, etc.).

[0031] Un signal de parole émis par un locuteur proche atteindra l'un des micros avant l'autre, et présentera donc un retard, et donc un déphasage φ, sensiblement constant. Pour du bruit, il peut certes exister également un déphasage entre les deux micros 10 et 12. En revanche, la notion de déphasage étant liée à la notion de direction d'onde incidente, on peut s'attendre à ce que ce déphasage soit différent de celui de la parole. Par exemple, si un bruit directif est dirigé dans le sens opposé à celui de la bouche, son déphasage sera de $-\varphi$ si le déphasage pour la voix est de φ. Dans le cas de l'invention, la réduction de bruit sur les signaux captés par les micros 10 et 12 n'est pas opérée dans le domaine

fréquentiel (comme cela est souvent le cas avec les techniques conventionnelles de débruitage) mais dans le domaine temporel.

[0032] Cette réduction de bruit est opérée au moyen d'un algorithme recherchant la fonction de transfert entre l'un des micros (par exemple le micro 10) et l'autre micro (le micro 12) au moyen d'un combineur adaptatif 14 mettant en oeuvre un filtre prédictif 16 de type LMS (*Least Mean Squares,* moindres carrés moyens). La sortie du filtre 16 est soustraite en 18 du signal du micro 10 pour donner un signal S débruité, appliqué en retour au filtre 16 pour permettre son adaptation itérative en fonction de l'erreur de prédiction. Il est ainsi possible de prédire à partir du signal capté par le micro 12 la composante de bruit contenue dans le signal capté par le micro 10 (la fonction de transfert identifiant le transfert du bruit).

[0033] La recherche adaptative de la fonction de transfert entre les deux micros n'est opérée que pendant les phases d'absence de parole. Pour cela, l'adaptation itérative du filtre 16 n'est activée que lorsqu'un détecteur 20 d'activité vocale VAD (*Voice Activity Detector*) piloté par un capteur 22 indique que le locuteur proche n'est pas en train de parler. Cette fonction est schématisée par le commutateur 24 : en l'absence de signal de parole avéré par le détecteur d'activité vocale 20, le combineur adaptatif 14 cherche à optimiser la fonction de transfert entre les deux micros 10 et 12 de manière à réduire la composante de bruit (position fermée du commutateur 24, comme illustré sur la figure) ; en revanche, en présence d'un signal de parole avéré par le détecteur d'activité vocale 20, le combineur adaptatif 14 fige les paramètres du filtre 16 à la valeur à laquelle ils se trouvaient juste avant que la parole ne soit détectée (ouverture du commutateur 24), ce qui évite toute dégradation du signal de parole du locuteur proche.

[0034] On notera que cette manière de procéder n'est pas gênante même en présence d'un environnement bruyant évolutif, car les mises à jour des paramètres du filtre 16 sont très fréquentes puisqu'elles interviennent à chaque fois que le locuteur proche cesse de parler.

[0035] De façon caractéristique de l'invention, le filtrage du combineur adaptatif 14 est un filtrage à délai fractionnaire, c'est-à-dire qu'il permet d'appliquer un filtrage entre les signaux captés par les deux micros en tenant compte d'un délai inférieur à la durée d'un échantillon de numérisation des signaux.

[0036] On sait qu'un signal temporel $x(t)$ de bande passante $[0,Fe/2]$ peut être reconstitué de manière parfaite à partir de la série discrète $x(k)$, où les échantillons $x(k)$ correspondent aux valeurs de $x(t)$ aux instants $k.Te$ ($Te = 1/Fe$ étant la période d'échantillonnage).

[0037] L'expression mathématique est la suivante :

$$x(t) = \sum_k x(k).\mathrm{sinc}\left(\frac{t - k.Te}{Te}\right)$$

[0038] La fonction sinus cardinal sinc étant définie par :

$$\mathrm{sinc}(t) = \frac{\sin(pi * t)}{pi * t}$$

[0039] La Figure 2 donne un représentation graphique de cette fonction sinc (t). Comme on peut le constater, cette fonction décroît rapidement, avec pour conséquence qu'un nombre fini et relativement faible de coefficients k dans la somme donne une très bonne approximation du résultat réel.

[0040] Pour un signal numérisé avec une période d'échantillonnage Te, l'intervalle ou décalage entre deux échantillons correspond de manière temporelle à une durée de Te seconde.

[0041] La série $x(n)$ des n échantillons successifs numérisés du signal capté peut ainsi être représentée par l'expression suivante, pour tout n entier :

$$x(n.Te) = \sum_k x(k).\mathrm{sinc}\left(\frac{n.Te - k.Te}{Te}\right)$$

[0042] On notera que dans la somme le terme en sinc est nul pour tout k, sauf pour $k = n$.

[0043] La Figure 3a donne un représentation graphique de cette fonction.

[0044] Si l'on veut calculer cette même série $x(n)$ décalée d'une valeur fractionnaire τ, c'est-à-dire d'un délai inférieur à la durée d'un échantillon de numérisation Te, l'expression ci-dessus devient :

$$x(n.Te-\tau)=\sum_k x(k).\mathrm{sinc}\left(\frac{(n-k).Te-\tau}{Te}\right)$$

[0045] La Figure 3b donne un représentation graphique de cette fonction, pour un exemple de valeur fractionnaire $\tau = 0,5$ (un demi-échantillon).

[0046] La série $x'(n)$ (décalée de τ) peut être vue comme la convolution de $x(n)$ par un filtre non causal G tel que :

$$x'(n)=G\otimes x(n)$$

[0047] Il s'agit donc de déterminer une estimée $\hat{G}$ d'un filtre optimal G tel que :

$$\hat{H}=\hat{G}\otimes\hat{F} \qquad \text{et} \qquad G(k)=\mathrm{sinc}(k+\tau/Te),$$

$\hat{H}$ étant l'estimée du transfert de bruit entre les deux micros, incluant un délai fractionnaire, et
F étant l'estimée de la réponse acoustique de l'environnement.

[0048] Pour l'estimation du filtre de transfert de bruit entre les deux micros, l'estimée $\hat{H}$ correspond à un filtre qui minimise une erreur :

$$e(n)=MicAvant(n)-\hat{H}*MicArrière(n)$$

[0049] *MicAvant(n)* et *MicArrière(n)* étant les valeurs respectives des signaux issus des capteurs microphoniques 10 et 12.

[0050] Ce filtre a pour caractéristique d'être non causal, c'est-à-dire qu'il se sert des échantillons futurs. En pratique, cela signifie que l'on introduit un retard dans le délai de traitement algorithmique. Comme il est non causal, il peut modéliser un délai fractionnaire et peut donc s'écrire $\hat{H}=\hat{G}\otimes\hat{F}$. (dans le cas classique d'un filtre causal, on aurait $\hat{H}=\hat{F}$).

[0051] Concrètement, dans l'algorithme, l'estimation de $\hat{H}$ a lieu directement, par la minimisation de l'erreur $e(n)$ ci-dessus, sans qu'il y a ait besoin d'estimer séparément $\hat{G}$ et $\hat{F}$.

[0052] Dans le cas classique causal (par exemple pour un filtre d'annulation d'écho), l'erreur $e(n)$ à minimiser s'écrit, sous forme développée :

$$e(n)=MicAvant(n)-\sum_{k=0}^{L-1}\hat{H}(k).MicArrière(n-k)$$

[0053] L étant la longueur du filtre.

[0054] Dans le cas de la présente invention (filtre non causal) l'erreur devient :

$$e(n) = MicAvant(n) - \sum_{k=-L}^{L-1} \hat{H}(k).MicArrière(n-k)$$

[0055] On notera que la longueur du filtre est doublée, pour tenir compte des échantillons futurs.

[0056] La prédiction du filtre *H* donne un filtre à délai fractionnaire qui, idéalement et en l'absence de parole, annule le bruit du micro 10 en ayant pour référence le micro 12 (comme on l'a indiqué plus haut, en période de parole le filtre est toutefois figé pour éviter toute dégradation de la parole locale).

[0057] Concrètement, le filtre $\hat{H}$ calculé par l'algorithme adaptatif qui estime le transfert de bruit entre le micro 10 et le micro 12, peut être vu comme la convolution $\hat{H} = \hat{G} \otimes \hat{F}$ de deux filtres $\hat{G}$ et $\hat{F}$ où :

- $\hat{G}$ correspond à la partie fractionnaire (avec la forme en sinus cardinal), et
- $\hat{F}$ correspond au transfert acoustique entre les deux micros, c'est-à-dire à la partie "environnementale" du système, représentative de l'acoustique du volume dans lequel opère celui-ci.

[0058] La Figure 4 illustre un exemple de réponse acoustique entre les deux micros, sous forme d'une caractéristique donnant l'amplitude *A* en fonction des coefficients *k* du filtre *F*. Les différentes réflexions du son qui peuvent intervenir en fonction de l'environnement, par exemple sur les vitres ou autres parois d'un habitacle de voiture, créent des pics visibles sur cette caractéristique de réponse acoustique.

[0059] La Figure 5 illustre un exemple du résultat de la convolution $G \otimes F$ des deux filtres *G* (réponse en sinus cardinal) et *F* (environnement d'utilisation), sous forme d'une caractéristique donnant l'amplitude A en fonction des coefficients *k* du filtre convolué.

[0060] L'estimée $\hat{H}$ peut être calculée par un algorithme LMS itératif cherchant à minimiser l'erreur $y(n) - \hat{H} \otimes x(n)$ pour converger vers le filtre optimal.

[0061] 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. Il s'agit d'algorithmes en eux-mêmes connus, décrits par exemple par :

[1] B. Widrow, Adaptative Filters, Aspect of Network and System Theory, R. E. Kalman and N. De Claris Eds., New York: Holt, Rinehart and Winston, pp. 563-587, 1970 ;

[2] B. Widrow et al., Adaptative Noise Cancelling: Principles and Applications, Proc. IEEE, Vol. 63, No 12 pp. 1692-1716, Dec 1975.

[3] B. Widrow et S. Stearns, Adaptative Signal Processing, Prentice-Hall Signal Processing Series, Alan V. Oppenheim Series Editor, 1985.

[0062] Comme on l'a indiqué plus haut, pour que le traitement précédent soit possible, il est nécessaire de disposer d'un détecteur d'activité vocale permettant de discriminer entre les phases d'absence de parole (où l'adaptation du filtre permet d'optimiser l'évaluation du bruit) et de présence de parole (où les paramètres du filtre sont figés à leur dernière valeur trouvée).

[0063] Plus précisément, le détecteur d'activité vocale est ici de préférence un détecteur "parfait", c'est-à-dire qu'il délivre un signal binaire (absence vs. présence de parole). Il se distingue ainsi de la plupart des détecteurs d'activité vocale utilisés dans les systèmes de débruitage connus, qui délivrent seulement une probabilité de présence de parole variable entre 0 et 100 % de façon continue ou par pas successifs. Avec de tels détecteurs basés seulement sur une probabilité de présence de parole, les fausses détections peuvent être importantes dans des environnements bruités. Pour être "parfait", le détecteur d'activité vocale ne peut pas se baser uniquement sur le signal capté par les micros ; il doit disposer d'une information additionnelle permettant de discriminer les phases de parole et de silence du locuteur proche.

[0064] Un premier exemple d'un tel détecteur est illustré par la Figure 6, où le détecteur d'activité vocale 20 opère en réponse au signal produit par une caméra.

[0065] Cette caméra est par exemple une caméra 26 installée dans l'habitacle d'un véhicule automobile, et orientée de manière que son angle de champ 28 englobe en toutes circonstances la tête 30 du conducteur, considéré comme le locuteur proche. Le signal délivré par la caméra 26 est analysé pour déterminer d'après le mouvement de la bouche et des lèvres si le locuteur parle ou non.

[0066] On peut utiliser à cet effet des algorithmes de détection de la région de la bouche dans une image d'un visage, et de suivi du mouvement des lèvres (*lip contour tracking*) telle que ceux exposés notamment par :

[4] G. Potamianos et al., Audio-Visual Automatic Speech Recognition: An Overview, Audio-Visual Speech Processing, G. Bailly et al. Eds., MIT Press, pp. 1-30, 2004.

[0067] Ce document décrit, de façon générale, l'apport d'une information visuelle en complément d'un signal audio pour notamment faire de la reconnaissance vocale dans des conditions acoustiques dégradées. Les données vidéo viennent ainsi s'ajouter aux données audio conventionnelles pour améliorer l'information vocale (*speech enhancement*).

[0068] Ce traitement pourra être utilisé dans le cadre de la présente invention pour distinguer entre les phases de parole et les phases de silence du locuteur. Pour tenir compte du fait que dans un habitacle automobile les mouvements de l'utilisateur sont lents tandis que les mouvements de la bouche sont rapides, on peut par exemple, une fois localisée la bouche, comparer deux images consécutives et évaluer le décalage sur un même pixel.

[0069] L'avantage de cette technique d'analyse d'image est de disposer d'une information complémentaire totalement indépendante de l'environnement de bruit acoustique.

[0070] Un autre exemple de capteur utilisable pour la détection d'activité vocale "parfaite" est un capteur physiologique susceptible de détecter certaines vibrations vocales du locuteur qui ne soient pas ou peu corrompues par le bruit environnant.

[0071] Un tel capteur peut être notamment constitué d'un accéléromètre ou d'un capteur piézoélectrique appliqué contre la joue ou la tempe du locuteur. En effet, lorsqu'une personne émet un son voisé (c'est-à-dire une composante de parole dont la production s'accompagne d'une vibration des cordes vocales), une vibration se propage depuis les cordes vocales jusqu'au pharynx et à la cavité bucco-nasale, où elle est modulée, amplifiée et articulée. La bouche, le voile du palais, le pharynx, les sinus et les fosses nasales servent ensuite de caisse de résonance à ce son voisé et, leur paroi étant élastique, elles vibrent à leur tour et ces vibrations sont transmises par conduction osseuse interne et sont perceptibles au niveau de la joue et de la tempe.

[0072] Ces vibrations au niveau de la joue et de la tempe présentent la caractéristique d'être, par nature, très peu corrompues par le bruit environnant : en effet, en présence de bruits extérieurs, même importants, les tissus de la joue et de la tempe ne vibrent quasiment pas, et ceci quelle que soit la composition spectrale du bruit extérieur.

[0073] Un capteur physiologique qui recueille ces vibrations vocales dépourvues de bruit donne un signal représentatif de la présence ou de l'absence de sons voisés émis par le locuteur, permettant donc de discriminer très bien les phases de parole et les phases de silence du locuteur.

[0074] Un tel capteur physiologique peut être notamment incorporé à un ensemble combiné micro/casque tel qu'illustré sur la Figure 7.

[0075] Sur cette figure, la référence 32 désigne de façon générale le casque selon l'invention, qui comporte deux oreillettes 34 réunies par un arceau. Chacune des oreillettes est de préférence constituée d'une coque fermée 36, logeant un transducteur de reproduction sonore, appliquée autour de l'oreille de l'utilisateur avec interposition d'un coussinet 38 isolant l'oreille de l'extérieur.

[0076] Le capteur physiologique 40 servant à la détection d'activité vocale est par exemple un accéléromètre intégré dans le coussinet 38 de manière à venir s'appliquer contre la joue ou la tempe de l'utilisateur avec un couplage le plus étroit possible. Ce capteur physiologique 40 peut notamment être placé sur la face intérieure de la peau du coussinet 38 de sorte que, une fois le casque mis en place, le capteur soit appliqué contre la joue ou la tempe de l'utilisateur sous l'effet d'une légère pression résultant de l'écrasement du matériau du coussinet, avec seulement interposition de la peau extérieure de ce coussinet.

[0077] Le casque porte également les micros 10, 12 du circuit de recueil et de débruitage de la parole du locuteur. Ces deux micros sont des micros omnidirectionnels placés sur la coque 36, et ils sont disposés avec le micro 10 placé en avant (plus proche de la bouche du porteur du casque) et le micro 12 placé plus en arrière. D'autre part la direction d'alignement 42 des deux micros 10, 12 est approximativement dirigée vers la bouche 44 du porteur du casque.

[0078] La Figure 8 est un schéma par blocs montrant les différentes fonctions mises en oeuvre par le combiné micro/casque de la Figure 7.

[0079] On retrouve sur cette figure les deux micros 10 et 12, ainsi que le détecteur d'activité vocale 20. Le micro avant 10 est le micro principal et le micro arrière 12 sert d'entrée au filtre adaptatif 16 du combineur 14. Le détecteur d'activité vocale 20 est contrôlé par le signal délivré par le capteur physiologique 40, avec par exemple lissage de la puissance du signal délivré par ce capteur 40 :

$$puissance_{capteur}(n) = \alpha . puissance_{capteur}(n-1) + (1-\alpha).(capteur(n))^2$$

α étant une constante de lissage proche de 1. Il suffit alors de fixer un seuil ζ tel que ce seuil soit dépassé dès que le locuteur parle.

[0080] La Figure 9 illustre l'allure des signaux recueillis :

- le signal S_{10} du chronogramme du haut correspond à ce qui est capté par le micro avant 10 : on voit qu'il est

impossible d'opérer à partir de ce signal (bruité) une discrimination efficace entre les phases de présence et d'absence de parole.

- le signal S_{40} du chronogramme du bas correspond à ce que délivre concurremment le capteur physiologique 40 : les phases successives de présence et d'absence de parole y sont marquées de façon bien plus apparente. Le signal binaire désigné VAD correspond à l'indication délivrée par le détecteur d'activité vocale 20 ('1' = présence de parole ; '0' = absence de parole), après évaluation de la puissance du signal S_{40} et comparaison par rapport au seuil ξ prédéfini.

[0081] Le signal délivré par le capteur physiologique 40 peut être utilisé non seulement comme signal d'entrée d'un détecteur d'activité vocale, mais également pour enrichir le signal capté par les micros 10 et 12, notamment dans le bas du spectre.

[0082] Bien sûr, les signaux délivrés par le capteur physiologique, qui correspondent aux sons voisés, ne sont pas à proprement parler de la parole puisque la parole n'est pas seulement formée de sons voisés, elle contient des composantes qui ne naissent pas au niveau des cordes vocales : le contenu fréquentiel est par exemple beaucoup plus riche avec le son provenant de la gorge et émis par la bouche. De plus, la conduction osseuse interne et la traversée de la peau a pour effet de filtrer certaines composantes vocales.

[0083] Par ailleurs, en raison du filtrage dû à la propagation des vibrations jusqu'à la tempe ou la joue, le signal recueilli par le capteur physiologique est utilisable uniquement dans les basses fréquences, principalement dans la région inférieure du spectre sonore (typiquement 0-1500 Hz).

[0084] Mais comme les bruits généralement rencontrés dans un environnement habituel (rue, métro, train, ...) sont majoritairement concentrés dans les basses fréquences, le signal d'un capteur physiologique présente l"avantage considérable d'être naturellement dépourvu de composante parasite de bruit il sera donc possible d'utiliser ce signal dans le bas du spectre, en le complétant dans le haut du spectre (au-dessus de 1500 Hz) par les signaux (bruités) recueillis par les micros 10 et 12, après avoir soumis ces signaux à une réduction de bruit opérée par le combineur adaptatif 14.

[0085] Le spectre complet est reconstruit au moyen du bloc de mixage 46 qui reçoit parallèlement : le signal du capteur physiologique 40 pour le bas du spectre, et le signal des micros 10 et 12 après débruitage par le combineur adaptatif 14 pour le haut du spectre. Cette reconstruction est opérée par sommation des signaux, qui sont appliqués en synchronisme au bloc de mixage 46 de manière à éviter toute déformation.

[0086] Le signal résultant délivré par le bloc 46 peut être soumis à une réduction de bruit finale par le circuit 48, opérée dans le domaine fréquentiel selon une technique conventionnelle comparable à celle décrite par exemple dans le WO 2007/099222 A1 (Parrot), pour donner en sortie le signal débruité final S.

[0087] La mise en oeuvre de cette technique est toutefois fortement simplifiée par rapport à ce qui est enseigné par exemple dans le document précité. En effet, dans le cas présent il n'est plus nécessaire d'évaluer une probabilité de présence de parole à partir du signal recueilli, puisque cette information peut être directement obtenue par le bloc de détection d'activité vocale 20 en réponse à la détection de l'émission de son voisé détecté par le capteur physiologique 40. L'algorithme peut être ainsi simplifié et rendu plus efficace et plus rapide.

[0088] La réduction de bruit fréquentielle est avantageusement opérée de façon différente en présence et en l'absence de parole (information donnée par le détecteur d'activité vocale parfait 20) :

- en l'absence de parole, la réduction de bruit est maximale sur toutes les bandes de fréquences, c'est-à-dire que le gain correspondant au débruitage maximum est appliqué de la même façon sur toutes les composantes du signal (puisque l'on est certain dans ce cas que celui-ci ne contient pas de composante utile) ;
- en revanche, en présence de parole, la réduction de bruit est une réduction fréquentielle, appliquée de façon différenciée sur chaque bande de fréquences selon le schéma classique.

[0089] Le système que l'on vient de décrire permet d'obtenir d'excellentes performances globales, typiquement de l'ordre de 30 à 40 dB de réduction de bruit sur le signal de parole du locuteur proche. Le combineur adaptatif 14 opérant sur les signaux captés par les micros 10 et 12 permet en particulier, avec le filtrage à délai fractionnaire, d'obtenir de très bonnes performances de débruitage dans les hautes fréquences.

[0090] Grâce à l'élimination de tous les bruits parasites, cela donne l'impression au locuteur distant (celui avec lequel le porteur du casque est en communication) que son interlocuteur (le porteur du casque) se trouve dans une pièce silencieuse.

Revendications

1. Un équipement audio, comprenant :

- un ensemble de deux capteurs microphoniques (10, 12) aptes à recueillir la parole de l'utilisateur de l'équipement et à délivrer des signaux de parole bruités respectifs ;
- des moyens d'échantillonnage des signaux de parole délivrés par les capteurs microphoniques ; et
- des moyens de débruitage d'un signal de parole, recevant en entrée les échantillons des signaux de parole délivrés par les deux capteurs microphoniques, et délivrant en sortie un signal de parole débruité représentatif de la parole émise par l'utilisateur de l'équipement,

dans lequel les moyens de débruitage sont des moyens de réduction de bruit non fréquentielle comprenant un combineur à filtre adaptatif (14) des signaux délivrés par les deux capteurs microphoniques, opérant par recherche itérative visant à annuler le bruit capté par l'un des capteurs microphoniques (10) sur la base d'une référence de bruit donnée par le signal délivré par l'autre capteur microphonique (12) ;
équipement **caractérisé en ce que** :

- le filtre adaptatif (16) est un filtre à délai fractionnaire, apte à modéliser un retard inférieur à la période d'échantillonnage des moyens d'échantillonnage ;
- l'équipement comprend en outre des moyens de détection d'activité vocale (20, 22) aptes à délivrer un signal représentatif de la présence ou de l'absence de parole par l'utilisateur de l'équipement, et
- le filtre adaptatif reçoit également en entrée le signal de présence ou d'absence de parole, de manière à, sélectivement : i) soit opérer une recherche adaptative des paramètres du filtre en l'absence de parole, ii) soit figer ces paramètres du filtre en présence de parole.

2. L'équipement audio de la revendication 1, dans lequel le filtre adaptatif (16) est apte à estimer un filtre optimal *H* tel que :

$$\hat{H} = \hat{G} \otimes \hat{F}$$

avec :

$$x'(n) = G \otimes x(n) \qquad \text{et} \qquad G(k) = \mathrm{sinc}(k + \tau / Te)\,,$$

$\hat{H}$ représentant l'estimée du filtre optimal *H*, transfert de bruit entre les deux capteurs microphoniques pour une réponse impulsionnelle incluant un délai fractionnaire,
$\hat{G}$ représentant l'estimée du filtre à délai fractionnaire *G* entre les deux capteurs microphoniques,
$\hat{F}$ représentant l'estimée de la réponse acoustique de l'environnement,
$\otimes$ indiquant une convolution,
x(*n*) étant la série d'échantillons du signal en entrée du filtre *H*,
x'(n) étant la série x(n) décalée d'un retard τ,
Te étant la période d'échantillonnage du signal en entrée du filtre *H*,
τ étant ledit délai fractionnaire, égal à un sous-multiple de *Te,* et
sinc indiquant la fonction sinus cardinal.

3. L'équipement audio de la revendication 1, dans lequel le filtre adaptatif est un filtre à algorithme de prédiction linéaire de type moindres carrés moyens LMS.

4. L'équipement audio de la revendication 1, dans lequel :

- l'équipement comprend en outre une caméra video (26) dirigée vers l'utilisateur (30) de l'équipement et apte à capter une image de celui-ci, et
- les moyens de détection d'activité vocale (20) comprennent des moyens d'analyse video aptes à analyser l'image produite par la caméra et à délivrer en réponse ledit signal de présence ou d'absence de parole par ledit utilisateur.

5. L'équipement audio de la revendication 1, dans lequel :

- l'équipement comprend en outre un capteur physiologique (40) apte à venir en contact avec la tête de l'utilisateur de l'équipement pour y être couplé afin de capter les vibrations vocales non acoustiques transmises par conduction osseuse interne, et
- les moyens de détection d'activité vocale (20) comprennent des moyens aptes à analyser le signal délivré par le capteur physiologique et à délivrer en réponse ledit signal de présence ou d'absence de parole par ledit utilisateur.

6. L'équipement audio de la revendication 5, dans lequel les moyens de détection d'activité vocale comprennent des moyens d'évaluation de l'énergie du signal délivré par le capteur physiologique, et des moyens à seuil.

7. L'équipement audio de la revendication 6, dans lequel l'équipement est un casque audio du type combiné micro/casque, comprenant :

- des écouteurs (34) comportant chacun un transducteur de reproduction sonore d'un signal audio logé dans une coque (36) pourvue d'un coussinet (38) circumaural ;
- lesdits deux capteurs microphoniques (10, 12), disposés sur la coque de l'un des écouteurs ; et
- ledit capteur physiologique (40), incorporé au coussinet de l'un des écouteurs et placé dans une région de celui-ci apte à venir en contact avec la joue ou la tempe du porteur du casque.

8. L'équipement audio de la revendication 7, dans lequel les deux capteurs microphoniques (10, 12) sont alignés en un réseau linéaire suivant une direction principale (42) dirigée vers la bouche (44) de l'utilisateur de l'équipement.

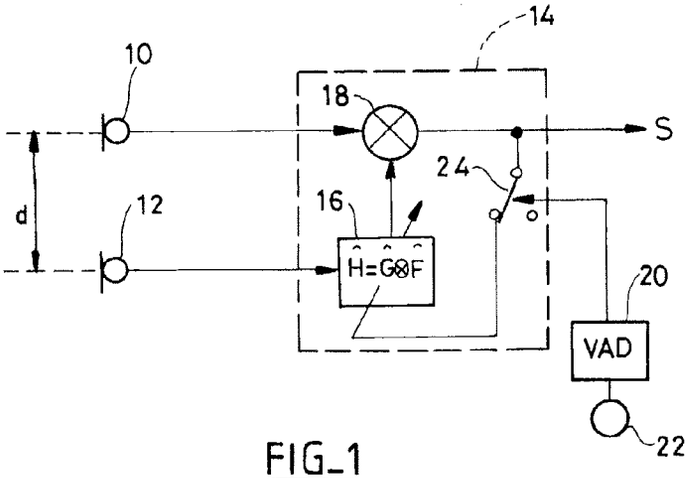

FIG_1

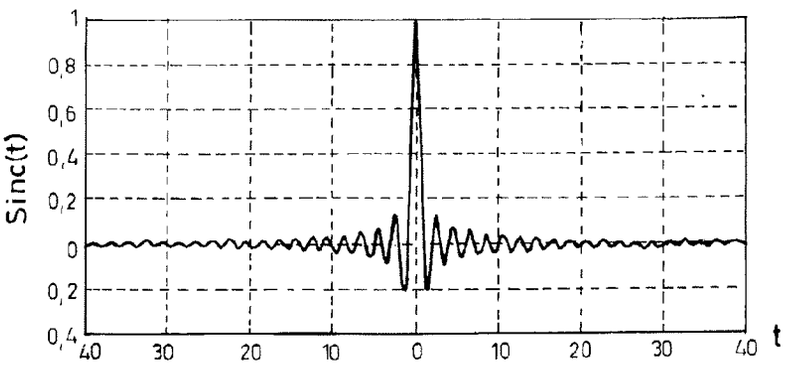

FIG_2

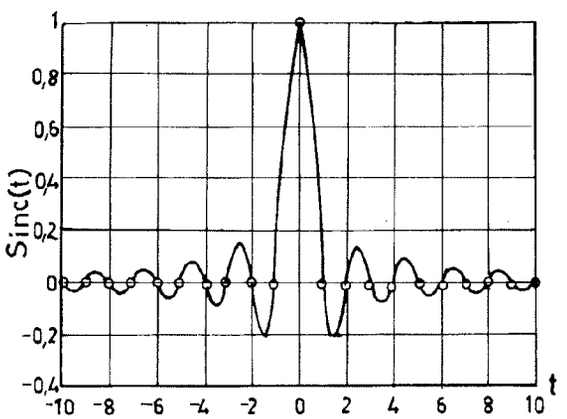

FIG_3a

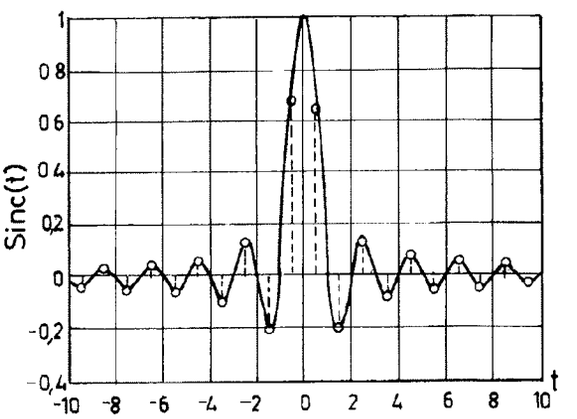

FIG_3b

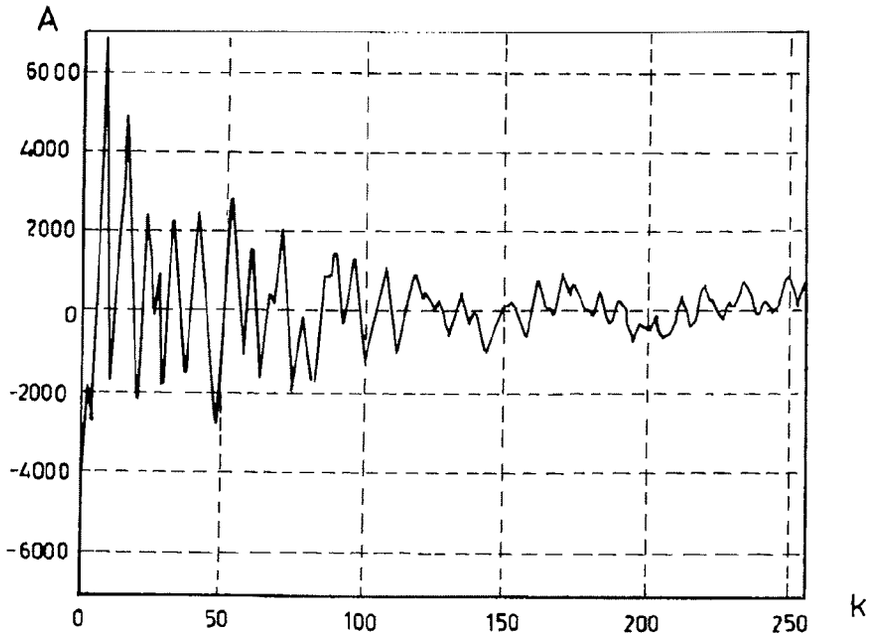
A
6000
4000
2000
0
-2000
-4000
-6000
0
50
100
150
200
250
k

FIG_4

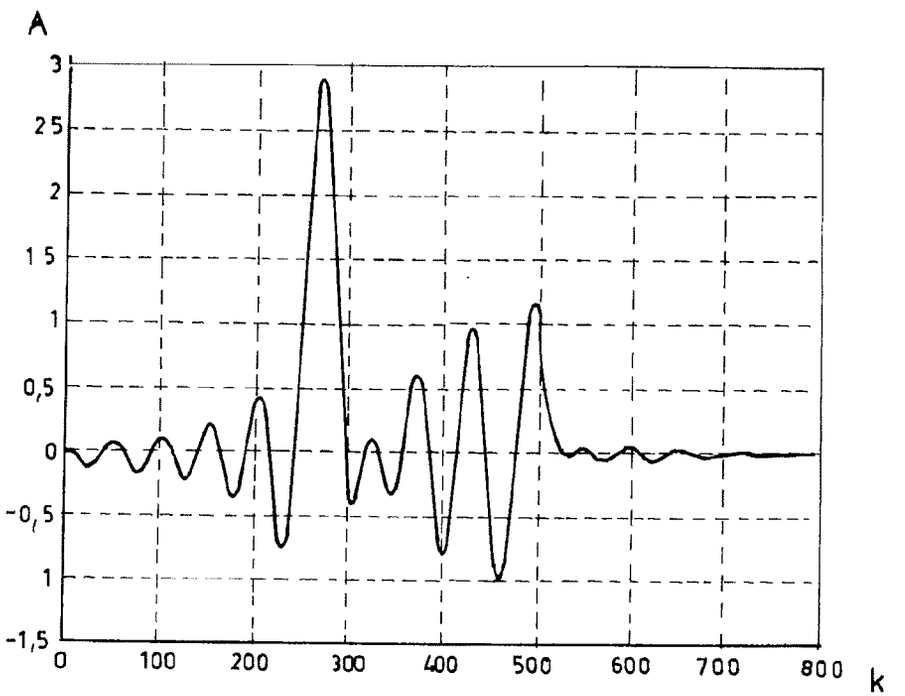
A
3
2,5
2
1,5
1
0,5
0
-0,5
-1
-1,5
0
100
200
300
400
500
600
700
800
k

FIG_5

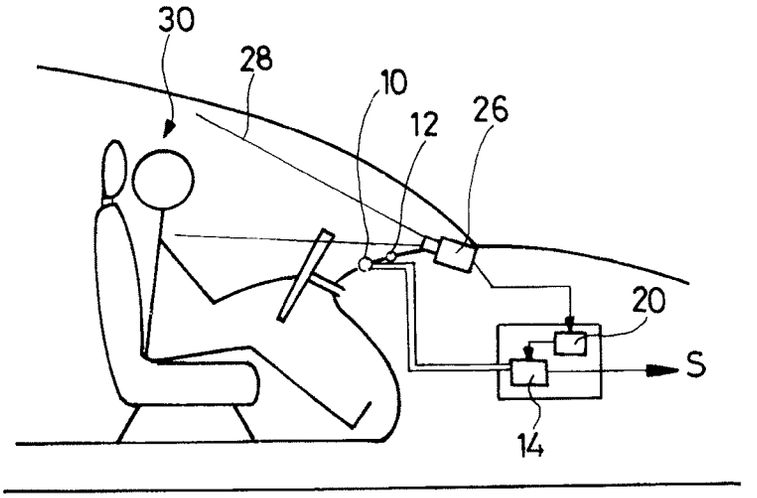

FIG_6

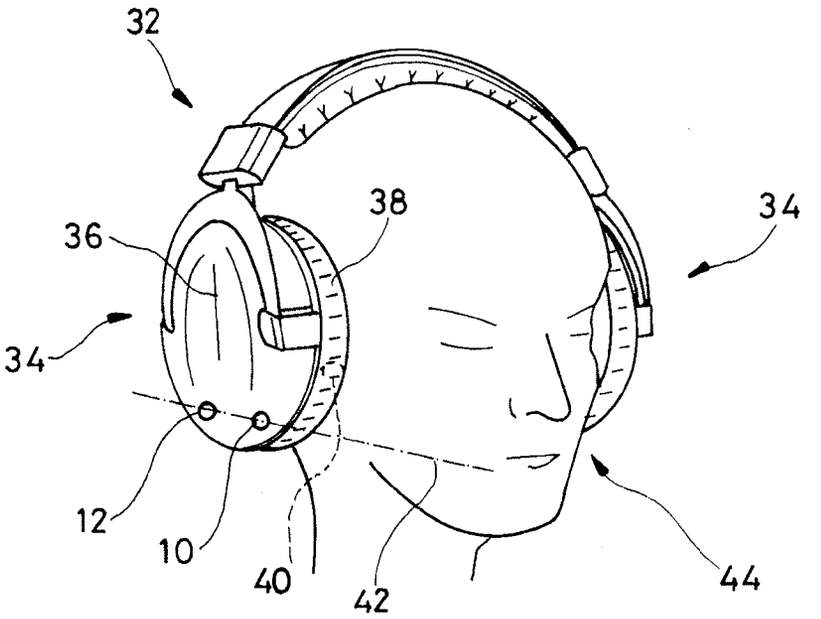

FIG_7

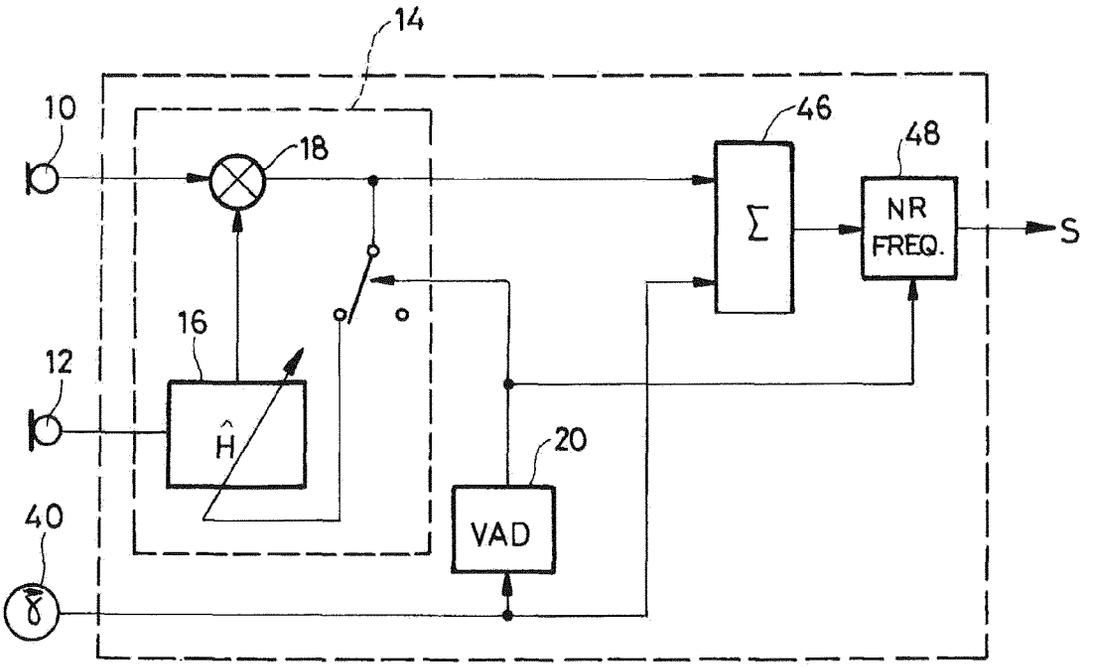

FIG_8

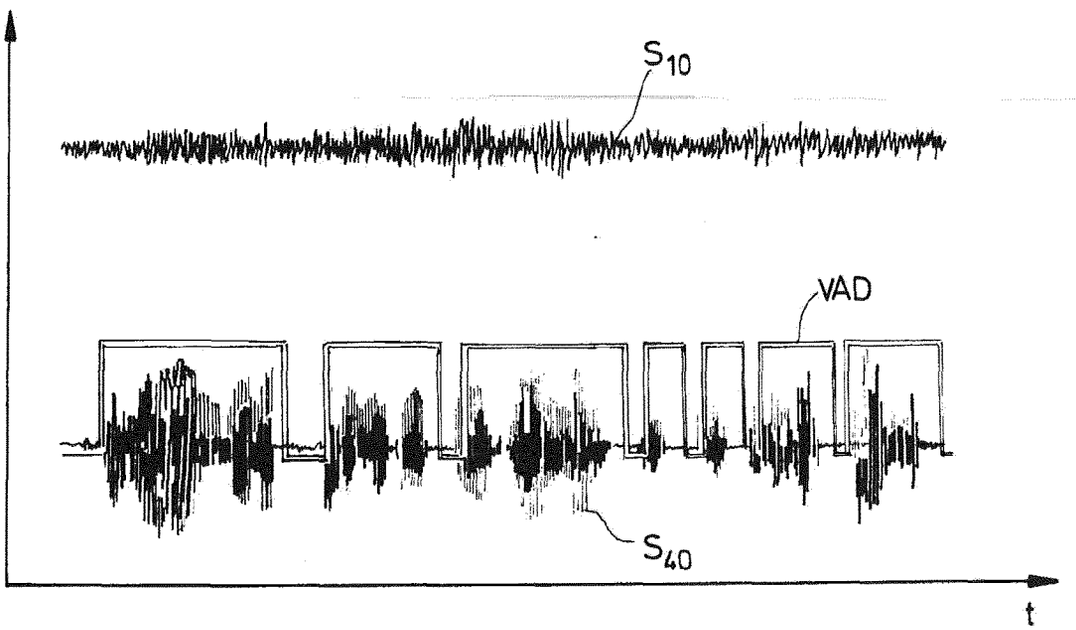

FIG_9

Europäisches Patentamt
European Patent Office
Office européen des brevets

RAPPORT DE RECHERCHE EUROPEENNE

Numéro de la demande
EP 12 17 0407

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	US 2008/280653 A1 (MA CHANGXUE [US] ET AL) 13 novembre 2008 (2008-11-13) * figure 4 * * alinéas [0036] - [0039] *	1	INV. G10L21/02
A	US 2007/165879 A1 (DENG HAO [CN] ET AL) 19 juillet 2007 (2007-07-19) * alinéas [0033], [0037], [0038], [0042] * * figure 4 * * alinéa [0049] *	1	
A	DJENDI M ET AL: "Noise Cancellation using Two Closely Spaced Microphones: Experimental Study witha Specific Model and Two Adaptive Algorithms", ACOUSTICS, SPEECH AND SIGNAL PROCESSING, 2006. ICASSP 2006 PROCEEDINGS . 2006 IEEE INTERNATIONAL CONFERENCE ON TOULOUSE, FRANCE 14-19 MAY 2006, PISCATAWAY, NJ, USA,IEEE, PISCATAWAY, NJ, USA, 14 mai 2006 (2006-05-14), page III, XP031386771, ISBN: 978-1-4244-0469-8 * abrégé * * alinéas [03.1], [0005] *	1	

DOMAINES TECHNIQUES RECHERCHES (IPC)

G10L
H04R

1 Le présent rapport a été établi pour toutes les revendications

Lieu de la recherche	Date d'achèvement de la recherche	Examinateur
Munich	22 juin 2012	Krembel, Luc

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

EPO FORM 1503 03.82 (P04C02)

ANNEXE AU RAPPORT DE RECHERCHE EUROPEENNE RELATIF A LA DEMANDE DE BREVET EUROPEEN NO. EP 12 17 0407

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.

22-06-2012

Document brevet cité au rapport de recherche	Date de publication	Membre(s) de la famille de brevet(s)	Date de publication
US 2008280653 A1	13-11-2008	US 2008280653 A1	13-11-2008
		WO 2008140931 A1	20-11-2008
US 2007165879 A1	19-07-2007	CN 1809105 A	26-07-2006
		US 2007165879 A1	19-07-2007

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] [0016]**
- US 20070165879 A1 **[0012]**
- WO 2007099222 A1, Parrot **[0086]**

Littérature non-brevet citée dans la description

- **B. WIDROW.** Adaptative Filters, Aspect of Network and System Theory. Holt, Rinehart and Winston, 1970, 563-587 **[0061]**
- **B. WIDROW et al.** Adaptative Noise Cancelling: Principles and Applications. *Proc. IEEE,* Décembre 1975, vol. 63 (12), 1692-1716 **[0061]**
- **B. WIDROW ; S. STEARNS.** Adaptative Signal Processing, Prentice-Hall Signal Processing Series. 1985 **[0061]**
- Audio-Visual Automatic Speech Recognition: An Overview. **G. POTAMIANOS et al.** Audio-Visual Speech Processing. MIT Press, 2004, 1-30 **[0066]**