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

(11) **EP 2 772 916 B1**

(12) **FASCICULE DE BREVET EUROPEEN**

(45) Date de publication et mention de la délivrance du brevet:
02.12.2015 Bulletin 2015/49

(51) Int Cl.:
G10L 21/0208 *(2013.01)* *G10L 25/18* *(2013.01)*
G10L 25/84 *(2013.01)*

(21) Numéro de dépôt: **14155968.2**

(22) Date de dépôt: **20.02.2014**

(54) **Procédé de débruitage d'un signal audio par un algorithme à gain spectral variable à dureté modulable dynamiquement**

Verfahren zur Geräuschdämpfung eines Audiosignals mit Hilfe eines Algorithmus mit variabler Spektralverstärkung mit dynamisch modulierbarer Härte

Method for suppressing noise in an audio signal by an algorithm with variable spectral gain with dynamically adaptive strength

(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

(30) Priorité: **28.02.2013 FR 1351760**

(43) Date de publication de la demande:
03.09.2014 Bulletin 2014/36

(73) Titulaire: **PARROT AUTOMOTIVE**
75010 Paris (FR)

(72) Inventeur: **Briot, Alexandre**
75010 Paris (FR)

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

(56) Documents cités:
US-A1- 2007 237 271 **US-B1- 7 454 010**

- **COHEN I ET AL: "Speech enhancement for non-stationary noise environments", SIGNAL PROCESSING, ELSEVIER SCIENCE PUBLISHERS B.V. AMSTERDAM, NL, vol. 81, no. 11, 1 novembre 2001 (2001-11-01), pages 2403-2418, XP004308517, ISSN: 0165-1684, DOI: 10.1016/S0165-1684(01)00128-1**
- **COHEN I: "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator", IEEE SIGNAL PROCESSING LETTERS, IEEE SERVICE CENTER, PISCATAWAY, NJ, US, vol. 9, no. 4, 1 avril 2002 (2002-04-01), pages 113-116, XP011433542, ISSN: 1070-9908, DOI: 10.1109/97.1001645**

EP 2 772 916 B1

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 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, 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) entraine la captation d'un niveau de parole relativement faible par rapport au bruit ambiant, qui rend difficile l'extraction du signal utile noyé dans le bruit. En plus de cette composante stationnaire permanente de bruit de roulement, 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 sélectif monocanal, c'est-à-dire opérant sur un unique signal (par opposition aux techniques mettant en oeuvre plusieurs micros dont les signaux sont combinés de façon judicieuse et font l'objet d'une analyse de cohérence spatiale ou spectrale, par exemple par des techniques de type *beamforming* ou autres). Cependant, elle s'appliquera avec la même pertinence à un signal recomposé à partir de plusieurs micros par une technique de *beamforming,* dans la mesure où l'invention présentée ici s'applique à un signal scalaire.

[0009] Dans le cas présent, il s'agit d'opérer le débruitage sélectif d'un signal audio bruité, généralement obtenu après numérisation du signal recueilli par un micro unique de l'équipement de téléphonie.

[0010] L'invention vise plus particulièrement un perfectionnement apporté aux algorithmes de réduction de bruit reposant sur un traitement du signal dans le domaine fréquentiel (donc après application d'une transformation de Fourier FFT) consistant à appliquer un gain spectral calculé en fonction de plusieurs estimateurs de probabilité de présence de parole.

[0011] Plus précisément, le signal y issu du microphone est découpé en trames de longueur fixe, chevauchantes ou non, et chaque trame d'indice k est transposée dans le domaine fréquentiel par FFT. Le signal fréquentiel résultant $Y(k, l)$, lui aussi discret, est alors décrit par un ensemble de "bins" fréquentiel (bandes de fréquences) d'indice l, typiquement 128 bins de fréquences positives.

[0012] Pour chaque trame de signal, un certain nombre d'estimateurs sont mis à jour pour déterminer une probabilité fréquentielle de présence de parole $p(k,l)$. Si la probabilité est grande, le signal sera considéré comme du signal utile (parole) et donc préservé avec un gain spectral $G(k,l) = 1$ pour le bin considéré. Dans le cas contraire, si la probabilité est faible le signal sera assimilé à du bruit et donc réduit, voire supprimé par application d'un gain spectral d'atténuation très inférieur à 1.

[0013] En d'autres termes, le principe de cet algorithme consiste à calculer et appliquer au signal utile un "masque fréquentiel" qui conserve l'information utile du signal de parole et élimine le signal parasite de bruit :

Cette technique peut être notamment implémentée par un algorithme de type OM-LSA (*Optimally Modified - Log*

Spectral Amplitude) telle que ceux décrits par :

[1] I. Cohen et B. Berdugo, "Speech Enhancement for Non-Stationary Noise Environments", Signal Processing, Vol. 81, No 11, pp. 2403-2418, Nov. 2001 ; et
[2] I. Cohen, "Optimal Speech Enhancement Under Signal Presence Uncertainty Using Log-Spectral Amplitude Estimator", IEEE Signal Processing Letters, Vol. 9, No 4, pp. 113-116, Apr. 2002.

[0014] Le US 7 454 010 B1 décrit également un algorithme comparable prenant en compte, pour le calcul des gains spectraux, une information de présence ou non de la voix dans un segment temporel courant.

[0015] 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.

[0016] L'efficacité d'une telle technique réside bien entendu dans le modèle de l'estimateur de probabilité de présence de parole qui doit discriminer parole et bruit.

[0017] Dans la pratique, l'implémentation d'un tel algorithme se heurte à un certain nombre de défauts, dont les deux principaux sont le "bruit musical" et l'apparition d'une "voix robotisée".

[0018] Le "bruit musical" se caractérise par une nappe de bruit de fond résiduel non uniforme, privilégiant certaines fréquences spécifiques. La tonalité du bruit n'est alors plus du tout naturelle, ce qui rend l'écoute perturbante. Ce phénomène résulte de ce que le traitement fréquentiel de débruitage est opéré sans dépendance entre fréquences voisines lors de la discrimination fréquentielle entre parole et bruit, car le traitement n'intègre pas de mécanisme pour prévenir deux gains spectraux voisins très différents. Or, dans les périodes de bruit seul, il faudrait idéalement un gain d'atténuation uniforme pour préserver la tonalité du bruit ; mais en pratique, si les gains spectraux ne sont pas homogènes, le bruit résiduel devient "musical" avec l'apparition de notes fréquentielles aux fréquences moins atténuées, correspondant à des bins faussement détectés comme contenant du signal utile. On notera que ce phénomène est d'autant plus marqué que l'on autorise l'application de gains d'atténuation importants.

[0019] Le phénomène de "voix robotisée" ou "voix métallique", quant à lui, se présente lorsque l'on choisit d'opérer une réduction de bruit très agressive, avec des gains spectraux d'atténuation importants. En présence de parole, des fréquences correspondant à de la parole mais qui sont faussement détectées comme étant du bruit seront fortement atténuées, rendant la voix moins naturelle, voire totalement artificielle ("robotisation" de la voix).

[0020] Le paramétrage d'un tel algorithme consiste donc à trouver un compromis sur l'agressivité du débruitage, de manière à enlever un maximum de bruit sans que les effets indésirables de l'application de gains spectraux d'atténuation trop importants ne deviennent trop perceptibles. Ce dernier critère se révèle toutefois extrêmement subjectif, et sur un groupe témoin d'utilisateurs relativement large il s'avère difficile de trouver un réglage de compromis qui puisse faire l'unanimité.

[0021] Pour minimiser ces défauts, inhérents à une technique de débruitage par application d'un gain spectral, le modèle "OM-LSA" prévoit de fixer une borne inférieure G_{min} pour le gain d'atténuation (exprimé suivant une échelle logarithmique, ce gain d'atténuation correspond donc dans la suite de ce document à une valeur négative) appliqué aux zones identifiées comme du bruit, de manière à s'interdire de trop débruiter pour limiter l'apparition des défauts évoqués plus haut. Cette solution n'est cependant pas optimale : certes, elle contribue à faire disparaitre les effets indésirables d'une réduction de bruit excessive, mais dans le même temps elle limite les performances du débruitage.

[0022] Le problème de l'invention est de pallier cette limitation, en rendant plus performant le système de réduction de bruit par application d'un gain spectral (typiquement selon un modèle OM-LSA), tout en respectant les contraintes évoquées plus haut, à savoir réduire efficacement le bruit sans altérer l'aspect naturel de la parole (en présence de parole) ni celui du bruit (en présence de bruit). En d'autres termes, il convient de rendre imperceptibles par le locuteur distant les effets indésirables du traitement algorithmique, tout en atténuant le bruit de manière importante.

[0023] L'idée de base de l'invention consiste à moduler le calcul du gain spectral G_{OMLSA} - calculé dans le domaine fréquentiel pour chaque bin - par un indicateur global, observé au niveau de la trame temporelle et non plus au niveau d'un unique bin de fréquence.

[0024] Cette modulation sera opérée par une transformation directe de la borne inférieure G_{min} du gain d'atténuation - borne qui est un scalaire communément désigné "dureté de débruitage" - en une fonction temporelle dont la valeur sera déterminée en fonction d'un descripteur temporel (ou "variable globale") reflété par l'état des divers estimateurs de l'algorithme. Ces derniers seront choisis en fonction de leur pertinence pour décrire des situations connues pour lesquelles on sait que le choix de la dureté de débruitage G_{min} peut être optimisé.

[0025] Par la suite et en fonction des cas de figure, la modulation temporelle appliquée à ce gain d'atténuation G_{min} logarithmique pourra correspondre soit à un incrément soit à un décrément : un décrément sera associé à une dureté de réduction de bruit plus grande (gain logarithmique plus grand en valeur absolue), inversement un incrément de ce gain logarithmique négatif sera associé à une valeur absolue plus petite donc une dureté de réduction de bruit plus faible.

[0026] En effet, on constate qu'une observation à l'échelle de la trame peut bien souvent permettre de corriger certains défauts de l'algorithme, notamment dans des zones très bruitées où il peut parfois faussement détecter une fréquence

de bruit comme étant une fréquence de parole : ainsi, si une trame de bruit seul est détectée (au niveau de la trame), on pourra débruiter de façon plus agressive sans pour autant introduire de bruit musical, grâce à un débruitage plus homogène.

[0027] Inversement, sur une période de parole bruitée, on pourra s'autoriser à moins débruiter afin de parfaitement préserver la voix tout en veillant à ce que la variation d'énergie du bruit de fond résiduel ne soit pas perceptible. On dispose ainsi d'un double levier (dureté et homogénéité) pour moduler l'importance du débruitage selon le cas considéré - phase de bruit seul ou bien phase de parole -, la discrimination entre l'un ou l'autre cas résultant d'une observation à l'échelle de la trame temporelle :

- dans le premier mode de réalisation, l'optimisation consistera à moduler dans le sens adéquat la valeur de la dureté de débruitage G_{min} pour mieux réduire le bruit en phase de bruit seul, et mieux préserver la voix en phase de parole ;

[0028] Plus précisément, l'invention propose un procédé de débruitage d'un signal audio par application d'un algorithme à gain spectral variable fonction d'une probabilité de présence de parole, comportant de manière en elle-même connue les étapes successives suivantes :

a) génération de trames temporelles successives du signal audio bruité numérisé ;
b) application d'une transformation de Fourier aux trames générées à l'étape a), de manière à produire pour chaque trame temporelle de signal un spectre de signal avec une pluralité de bandes de fréquences prédéterminées ;
c) dans le domaine fréquentiel :

c1) estimation, pour chaque bande de fréquences de chaque trame temporelle courante, d'une probabilité de présence de parole ;
c3) calcul d'un gain spectral, propre à chaque bande de fréquence de chaque trame temporelle courante, en fonction de : i) une estimation de l'énergie du bruit dans chaque bande de fréquences, ii) la probabilité de présence de parole estimée à l'étape c1), et iii) une valeur scalaire de gain minimal représentative d'un paramètre de dureté du débruitage ;
c4) réduction sélective de bruit par application à chaque bande de fréquences du gain calculé à l'étape c3) ;

d) application d'une transformation de Fourier inverse au spectre de signal constitué des bandes de fréquences produites à l'étape c4), de manière à délivrer pour chaque spectre une trame temporelle de signal débruité ; et
e) reconstitution d'un signal audio débruité à partir des trames temporelles délivrées à l'étape d).

[0029] De façon caractéristique de l'invention :

- ladite valeur scalaire de gain minimal est une valeur modulable de manière dynamique à chaque trame temporelle successive ; et
- le procédé comporte en outre, préalablement à l'étape c3) de calcul du gain spectral, une étape de :

c2) calcul, pour la trame temporelle courante, de ladite valeur modulable en fonction d'une variable globale observée au niveau de la trame temporelle courante pour toutes les bandes de fréquences ; et

- ledit calcul de l'étape c2) comprend l'application, pour la trame temporelle courante, d'un incrément/décrément apporté à une valeur paramétrée nominale dudit gain minimal.

[0030] Dans une première implémentation de l'invention, la variable globale est un rapport signal sur bruit de la trame temporelle courante, évalué dans le domaine temporel.

[0031] La valeur scalaire de gain minimal peut notamment être calculée à l'étape c2) par application de la relation :

$$G_{min}(k) = G_{min} + \Delta G_{min}(SNR_y(k))$$

k étant l'indice de la trame temporelle courante,
$G_{min}(k)$ étant le gain minimal à appliquer à la trame temporelle courante,
G_{min} étant ladite valeur nominale paramétrée du gain minimal,
$\Delta G_{min}\ (k)$ étant ledit incrément/décrément apporté à G_{min}, et
$SNR_y\ (k)$ étant le rapport signal sur bruit de la trame temporelle courante.

[0032] Dans une deuxième implémentation de l'invention, la variable globale est une probabilité moyenne de parole, évaluée au niveau de la trame temporelle courante.

[0033] La valeur scalaire de gain minimal peut notamment être calculée à l'étape c2) par application de la relation :

$$G_{min}(k) = G_{min} + \left(P_{speech}(k) - 1\right).\Delta_1 G_{min} + P_{speech}(k).\Delta_2 G_{min}$$

k étant l'indice de la trame temporelle courante,
$G_{min}(k)$ étant le gain minimal à appliquer à la trame temporelle courante,
G_{min} étant ladite valeur nominale paramétrée du gain minimal,
$P_{speech}(k)$ étant la probabilité moyenne de parole évaluée au niveau de la trame temporelle courante,
$\Delta_1 G_{min}$ étant ledit incrément/décrément, apporté à G_{min} en phase de bruit, et
$\Delta_2 G_{min}$ étant ledit incrément/décrément, apporté à G_{min} en phase de parole.

[0034] La probabilité moyenne de parole peut notamment être évaluée au niveau de la trame temporelle courante par application de la relation :

$$P_{speech}(k) = \frac{1}{N}\sum_{l}^{N} p(k,l)$$

l étant l'indice de la bande de fréquences,
N étant le nombre de bandes de fréquences dans le spectre, et
$p(k,l)$ étant la probabilité de présence de parole de la bande de fréquences d'indice *l* de la trame temporelle courante.

[0035] Dans une troisième implémentation de l'invention, la variable globale est un signal booléen de détection d'activité vocale pour la trame temporelle courante, évalué dans le domaine temporel par analyse de la trame temporelle et/ou au moyen d'un détecteur externe.

[0036] La valeur scalaire de gain minimal peut notamment être calculée à l'étape c2) par application de la relation :

$$G_{min}(k) = G_{min} + VAD(k).\Delta G_{min}$$

k étant l'indice de la trame temporelle courante,
$G_{min}(k)$ étant le gain minimal à appliquer à la trame temporelle courante,
G_{min} étant ladite valeur nominale paramétrée du gain minimal,
VAD (k) étant la valeur du signal booléen de détection d'activité vocale pour la trame temporelle courante, et
ΔG_{min} étant ledit incrément/décrément apporté à G_{min}.

[0037] 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 un traitement de débruitage de type OM-LSA selon l'état de la technique.
La Figure 2 illustre le perfectionnement apporté par l'invention à la technique de débruitage de la Figure 1.

[0038] Le processus de l'invention est mis en oeuvre par des moyens logiciels, schématisés sur les figures 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.

Algorithme de débruitage OM-LSA selon l'état de la technique

[0039] La Figure 1 illustre de façon schématique, sous forme de blocs fonctionnels, la manière dont est réalisé un traitement de débruitage de type OM-LSA selon l'état de la technique.

[0040] Le signal numérisé $y(n) = x(n) + d(n)$ comprenant une composante de parole $x(n)$ et une composante de bruit $d(n)$ (n étant le rang de l'échantillon) est découpé (bloc 10) en segments ou trames temporelles $y(k)$ (k étant l'indice de la trame) de longueur fixe, chevauchantes ou non, habituellement des trames de 256 échantillons pour un signal échantillonné à 8 kHz (standard téléphonique *narrowband*).

[0041] Chaque trame temporelle d'indice k est ensuite transposée dans le domaine fréquentiel par une transformation rapide de Fourier FFT (bloc 12) : le signal résultant obtenu ou spectre $Y(k,l)$, lui aussi discret, est alors décrit par un ensemble de bandes de fréquences ou "bins" fréquentiels (l étant l'indice de bin), par exemple 128 bins de fréquences positives. Un gain spectral $G = G_{OMLSA}(k,l)$, propre à chaque bin, est appliqué (bloc 14) au signal fréquentiel $Y(k,l)$, pour donner un signal $\hat{X}(k, l)$:

$$\hat{X}(k,l) = G_{OMLSA}(k,l).\ Y(k,l)$$

[0042] Le gain spectral $G_{OMLSA}(k,l)$ est calculé (bloc 16) en fonction d'une part d'une probabilité de présence de parole $p(k,l)$, qui est une probabilité fréquentielle évaluée (bloc 18) pour chaque bin, et d'autre part d'un paramètre G_{min}, qui est une valeur scalaire de gain minimal, dénommée couramment "dureté de débruitage". Ce paramètre G_{min} fixe une borne inférieure au gain d'atténuation appliqué sur les zones identifiées comme du bruit, afin d'éviter que les phénomènes de bruit musical et de voix robotisée ne deviennent trop marqués du fait de l'application de gains spectraux d'atténuation trop importants et/ou hétérogènes.

[0043] Le gain spectral $G_{OMLSA}(k,l)$ calculé est de la forme :

$$G_{OMLSA}(k,l) = \{G(k,l)\}^{p(k,l)}.\ G_{min}^{1-p(k,l)}$$

[0044] Le calcul du gain spectral et celui de la probabilité de présence de parole sont donc avantageusement implémentés sous forme d'un algorithme de type OM-LSA *(Optimally Modified - Log Spectral Amplitude)* tel que celui décrit dans l'article (précité) :

> [2] I. Cohen, "Optimal Speech Enhancement Under Signal Presence Uncertainty Using Log-Spectral Amplitude Estimator", IEEE Signal Processing Letters, Vol. 9, No 4, pp. 113-116, Apr. 2002.

[0045] 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 critère se montre adapté, car la distance choisie est en meilleure adéquation avec le comportement de l'oreille humaine et donne donc qualitativement de meilleurs résultats.

[0046] 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.

[0047] L'algorithme "OM-LSA" (*Optimally-Modified LSA*) améliore le calcul du gain LSA en le pondérant par la probabilité conditionnelle $p(k,l)$ de présence de parole ou SPP (*Speech Presence Probability*), 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.

[0048] La probabilité de présence de parole $p(k,l)$ 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 :

> [3] 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.

[0049] Comme fréquemment dans ce domaine, le procédé décrit n'a pas pour objectif d'identifier précisément sur quelles composantes fréquentielles de quelles trames la parole est absente, mais plutôt de donner un indice de confiance entre 0 et 1, une valeur 1 indiquant que la parole est absente à coup sûr (selon l'algorithme) tandis qu'une valeur 0

déclare le contraire. Par sa nature, cet indice est assimilé à la probabilité d'absence de la parole *a priori,* c'est-à-dire la probabilité que la parole soit absente sur une composante fréquentielle donnée de la trame considérée. Il s'agit bien sûr d'une assimilation non rigoureuse, dans le sens que même si la présence de la parole est probabiliste *ex ante,* le signal capté par le micro ne présente à chaque instant que l'un de deux états distincts : à l'instant considéré, il peut soit comporter de la parole soit ne pas en contenir. En pratique, cette assimilation donne toutefois de bons résultats, ce qui justifie son utilisation.

[0050] On pourra également se référer au WO 2007/099222 A1 (Parrot), qui décrit en détail une technique de débruitage dérivée de ce principe, mettant en oeuvre un calcul de probabilité de présence de parole.

[0051] Le signal résultant $\hat{X}(k,l) = G_{OMLSA}(k,l).\ Y(k,l)$, c'est-à-dire le signal utile $Y(k,l)$ auquel a été appliqué le masque fréquentiel $G_{OMLSA}(k,l)$, fait ensuite l'objet d'une transformation de Fourier inverse iFFT (bloc 20), pour repasser du domaine fréquentiel au domaine temporel. Les trames temporelles obtenues sont ensuite rassemblées (bloc 22) pour donner un signal débruité numérisé $\hat{x}(n)$.

Algorithme de débruitage OM-LSA selon l'invention

[0052] La Figure 2 illustre les modifications apportées à l'algorithme que l'on vient d'exposer. Les blocs portant les mêmes références numériques correspondent à des fonctions identiques ou similaires à celles exposées plus haut, de même que les références des divers signaux traités.

[0053] Dans l'implémentation connue de la Figure 1, la valeur scalaire G_{min} du gain minimal représentatif de la dureté de débruitage était choisie plus ou moins empiriquement, de telle sorte que la dégradation de la voix reste peu audible, tout en assurant une atténuation acceptable du bruit. Comme on l'a exposé en introduction, il est cependant souhaitable de débruiter plus agressivement en phase de bruit seul, mais sans pour autant introduire de bruit musical ; inversement, sur une période de parole bruitée, on peut s'autoriser à moins débruiter afin de parfaitement préserver la voix tout en veillant à ce que la variation d'énergie du bruit de fond résiduel ne soit pas perceptible.

[0054] On peut disposer selon le cas (phase de bruit seul ou bien phase de parole) d'un double intérêt à moduler la dureté du débruitage : celle-ci sera modulée en faisant varier dynamiquement la valeur scalaire de G_{min}, dans le sens adéquat qui réduira le bruit en phase de bruit seul et préservera mieux la voix en phase de parole.

[0055] Pour ce faire, la valeur scalaire G_{min}, initialement constante, est transformée (bloc 24) en une fonction temporelle $G_{min}(k)$ dont la valeur sera déterminée en fonction d'une variable globale (également désignée "descripteur temporel"), c'est-à-dire d'une variable considérée globalement au niveau de la trame et non pas du bin fréquentiel. Cette variable globale peut être reflétée par l'état d'un ou plusieurs estimateurs différents déjà calculés par l'algorithme, qui seront choisis selon le cas en fonction de leur pertinence.

[0056] Ces estimateurs peuvent notamment être : i) un rapport signal sur bruit, ii) une probabilité moyenne de présence de parole et/ou iii) une détection d'activité vocale. Dans tous ces exemples, la dureté de débruitage G_{min} devient une fonction temporelle $G_{min}(k)$ définie par les estimateurs, eux-mêmes temporels, permettant de décrire des situations connues pour lesquelles on souhaite moduler la valeur de G_{min} afin d'influer sur la réduction de bruit en modifiant de façon dynamique le compromis débruitage/dégradation du signal.

[0057] On notera incidemment que, pour que cette modulation dynamique de la dureté ne soit pas perceptible par l'auditeur, il convient de prévoir un mécanisme pour prévenir des variations brutales de $G_{min}(k)$, par exemple par une technique conventionnelle de lissage temporel. On évitera ainsi que des variations temporelles brusques de la dureté $G_{min}(k)$ ne soient audibles sur le bruit résiduel, qui est très souvent stationnaire dans le cas par exemple d'un automobiliste en condition de roulage.

Descripteur temporel : rapport signal sur bruit

[0058] Le point de départ de cette première implémentation est la constatation de ce qu'un signal de parole capté dans un environnement silencieux n'a que peu, voire pas, besoin d'être débruité, et qu'un débruitage énergique appliqué à un tel signal conduirait rapidement à des artefacts audibles, sans que le confort d'écoute ne soit amélioré du seul point de vue du bruit résiduel.

[0059] À l'inverse, un signal excessivement bruité peu rapidement devenir inintelligible ou susciter une fatigue progressive à l'écoute ; dans un tel cas le bénéfice d'un débruitage important sera indiscutable, même au prix d'une dégradation audible (toutefois raisonnable et contrôlée) de la parole.

[0060] En d'autres termes, la réduction de bruit sera d'autant plus bénéfique pour la compréhension du signal utile que le signal non traité est bruité.

[0061] Ceci peut être pris en compte en modulant le paramètre de dureté G_{min} en fonction du rapport signal sur bruit a priori ou du niveau de bruit courant du signal traité :

$$G_{min}(k) = G_{min} + \Delta G_{min}(SNR_y(k))$$

$G_{min}(k)$ étant le gain minimal à appliquer à la trame temporelle courante,
G_{min} étant une valeur nominale paramétrée de ce gain minimal,
$\Delta G_{min}(k)$ étant l'incrément/décrément apporté à la valeur G_{min}, et
$SNR_y(k)$ étant le rapport signal sur bruit de la trame courante, évalué dans le domaine temporel (bloc 26), correspondant à la variable appliquée sur l'entrée n° ① du bloc 24 (ces "entrées" étant symboliques et n'ayant qu'une valeur illustrative des différentes possibilités alternatives de mise en oeuvre de l'invention).

Descripteur temporel : probabilité moyenne de présence de parole

[0062] Un autre critère pertinent pour moduler la dureté de la réduction peut être la présence de parole pour la trame temporelle considérée.

[0063] Avec l'algorithme conventionnel, lorsqu'on tente d'augmenter la dureté de débruitage G_{min}, le phénomène de "voix robotisée" apparait avant celui de "bruit musical". Ainsi, il parait possible et intéressant d'appliquer une dureté de débruitage plus grande dans une phase de bruit seul, en modulant simplement le paramètre de dureté de débruitage par un indicateur global de présence de parole : en période de bruit seul, le bruit résiduel - à l'origine de la fatigue d'écoute - sera réduit par application d'une dureté plus importante, et ce sans contrepartie puisque la dureté en phase de parole peut rester inchangée.

[0064] Comme l'algorithme de réduction de bruit repose sur un calcul de probabilité de présence de parole fréquentielle, il est aisé d'obtenir un indice moyen de présence de parole à l'échelle de la trame à partir des différentes probabilités fréquentielles, de manière à différencier les trames principalement constituées de bruit de celles qui contiennent de la parole utile. On peut par exemple utiliser l'estimateur classique :

$$P_{speech}(k) = \frac{1}{N}\sum_{l}^{N} p(k,l)$$

$P_{speech}(k)$ étant la probabilité moyenne de parole évaluée au niveau de la trame temporelle courante,
N étant le nombre de bins du spectre, et
$p(k,l)$ étant la probabilité de présence de parole du bin d'indice l de la trame temporelle courante.

[0065] Cette variable $P_{speech}(k)$ est calculée par le bloc 28 et appliquée sur l'entrée n° ② du bloc 24, qui calcule la dureté de débruitage à appliquer pour une trame donnée :

$$G_{min}(k) = G_{min} + \left(P_{speech}(k) - 1\right).\Delta_1 G_{min} + P_{speech}(k).\Delta_2 G_{min}$$

$G_{min}(k)$ étant le gain minimal à appliquer à la trame temporelle courante,
G_{min} étant une valeur nominale paramétrée de ce gain minimal, et
$\Delta_1 G_{min}$ étant un incrément/décrément apporté à G_{min} en phase de bruit, et
$\Delta_2 G_{min}$ étant un incrément/décrément apporté à G_{min} en phase de parole.

[0066] L'expression ci-dessus met bien en évidence les deux effets complémentaires de l'optimisation présentée, à savoir :

- l'augmentation de la dureté de la réduction de bruit d'un facteur $\Delta_1 G_{min}$ en phase de bruit afin de réduire le bruit résiduel, typiquement $\Delta_1 > 0$, par exemple Δ_1 = +6 dB ; et
- la diminution de la dureté de la réduction de bruit d'un facteur $\Delta_2 G_{min}$ en phase de parole afin de mieux préserver la voix, typiquement $\Delta_2 < 0$, par exemple Δ_2 = -3 dB.

Descripteur temporel : détecteur d'activité vocale

[0067] Dans cette troisième implémentation, un détecteur d'activité vocale ou VAD (bloc 30) est mis à profit pour

effectuer le même type de modulation de dureté que dans l'exemple précédent. Un tel détecteur "parfait" délivre un signal binaire (absence vs. présence de parole), et se distingue des systèmes délivrant seulement une probabilité de présence de parole variable entre 0 et 100 % de façon continue ou par pas successifs, qui peuvent introduire des fausses détections importantes dans des environnements bruités.

[0068] Le module de détection d'activité vocale ne prenant que deux valeurs distinctes '0' ou '1', la modulation de la dureté de débruitage sera discrète :

$$G_{min}(k) = G_{min} + VAD(k).\Delta G_{min}$$

$G_{min}(k)$ étant le gain minimal à appliquer à la trame temporelle courante,
G_{min} étant une valeur nominale paramétrée dudit gain minimal,
VAD (k) étant la valeur du signal booléen de détection d'activité vocale pour la trame temporelle courante, évalué dans le domaine temporel (bloc 30) et appliqué à l'entrée n° ③ du bloc 24, et
ΔG_{min} étant l'incrément/décrément apporté à la valeur G_{min}.

[0069] Le détecteur d'activité vocale 30 peut être réalisé de différentes manières, dont a va donner ci-dessous trois exemples d'implémentation.

[0070] Dans un *premier exemple,* la détection est opérée à partir du signal $y(k)$, d'une manière intrinsèque au signal recueilli par le micro ; une analyse du caractère plus ou moins harmonique de ce signal permet de déterminer la présence d'une activité vocale, car un signal présentant une forte harmonicité peut être considéré, avec une faible marge d'erreur, comme étant un signal de voix, donc correspondant à une présence de parole.

[0071] Dans un *deuxième exemple,* le détecteur d'activité vocale 30 fonctionne en réponse au signal produit par une caméra, installée par exemple dans l'habitacle d'un véhicule automobile et orientée de manière que son angle de champ englobe en toutes circonstances la tête du conducteur, considéré comme le locuteur proche. Le signal délivré par la caméra est analysé pour déterminer d'après le mouvement de la bouche et des lèvres si le locuteur parle ou non, comme cela est décrit entre autres dans le EP 2 530 672 A1 (Parrot SA), auquel on pourra se référer pour de plus amples explications. L'avantage de cette technique d'analyse d'image est de disposer d'une information complémentaire totalement indépendante de l'environnement de bruit acoustique.

[0072] Un *troisième exemple* de capteur utilisable pour la détection d'activité vocale est un capteur physiologique susceptible de détecter certaines vibrations vocales du locuteur qui ne sont pas ou peu corrompues par le bruit environnant. Un tel capteur peut être notamment constitué d'un accéléromètre ou d'un capteur piézoélectrique appliqué contre la joue ou la tempe du locuteur. Il peut être en particulier incorporé au coussinet d'un écouteur d'un ensemble combiné micro/casque, comme cela est décrit dans le EP 2 518 724 A1 (Parrot SA), auquel on pourra se reporter pour plus de détails.

[0073] 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.

[0074] 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. 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.

Variante de mise en oeuvre de l'algorithme de débruitage OM-LSA

[0075] En variante ou en complément de ce qui précède, le gain spectral G_{OMLSA} - calculé dans le domaine fréquentiel pour chaque bin - peut être modulé de façon indirecte, en pondérant la probabilité de présence de parole fréquentielle $p(k,l)$ par un indicateur global temporel observé au niveau de la trame (et non plus d'un simple bin fréquentiel particulier).

[0076] Dans ce cas, si une trame de bruit seul est détectée, on peut avantageusement considérer que chaque probabilité fréquentielle de parole devrait être nulle, et la probabilité fréquentielle locale pourra être pondérée par une donnée globale, cette donnée globale permettant de faire une déduction sur le cas réel rencontré à l'échelle de la trame (phase de parole/phase de bruit seul), que la seule donnée dans le domaine fréquentiel n'autorise pas à formuler ; en

présence de bruit seul, on pourra se ramener à un débruitage uniforme, évitant toute musicalité du bruit, qui gardera son "grain" d'origine.

[0077] En d'autres termes, la probabilité de présence de parole initialement fréquentielle sera pondérée par une probabilité de présence globale de parole à l'échelle de la trame : on s'efforcera alors de débruiter de manière homogène l'ensemble de la trame dans un cas d'absence de parole (débruiter uniformément quand la parole est absente).

[0078] En effet, comme on l'a exposé plus haut, de présence de parole $P_{speech}(k)$ (calculée comme la moyenne arithmétique des probabilités fréquentielles de présence de parole) est un indicateur plutôt fiable de la présence de parole à l'échelle de la trame. On peut alors envisager de modifier l'expression conventionnelle du calcul du gain OM-LSA, à savoir :

$$G_{OMLSA}(k,l) = \{G(k,l)\}^{p(k,l)} . G_{min}^{1-p(k,l)}$$

en pondérant la probabilité fréquentielle de présence de parole par une donnée globale $p_{glob}(k)$ de présence de parole évaluée au niveau de la trame :

$$G_{OMLSA}(k,l) = \{G(k,l)\}^{p(k,l).p_{glob}(k)} . G_{min}^{1-p(k,l).p_{glob}(k)}$$

$G_{Oivf_{LSA}}$ (k,l) étant le gain spectral à appliquer au bin d'indice l de la trame temporelle courante,
$G(k,l)$ étant un gain de débruitage sous-optimal à appliquer au bin d'indice l,
$p(k,l)$ étant la probabilité de présence de parole du bin d'indice l de la trame temporelle courante,
$p_{glob}(k)$ étant la probabilité globale et seuillée de parole, évaluée au niveau de la trame temporelle courante, et
G_{min} étant une valeur nominale paramétrée du gain spectral.

[0079] La donnée globale $p_{glob}(k)$ au niveau de la trame temporelle peut notamment être évaluée de la manière suivante :

$$p_{glob}(k) = \frac{1}{P_{seuil}} . max\left\{P_{speech}(k); P_{seuil}\right\}$$

$$P_{speech}(k) = \frac{1}{N}\sum_{l}^{N} p(k,l)$$

P_{seuil} étant une valeur de seuil de la probabilité globale de parole, et
N étant le nombre de bins dans le spectre.

[0080] Ceci revient à substituer dans l'expression conventionnelle la probabilité fréquentielle $p(k,l)$ par une probabilité combinée $p_{combinée}(k,l)$ qui intègre une pondération par la donnée globale $p_{glob}(k)$, non fréquentielle, évaluée au niveau de la trame temporelle en présence de parole :

$$G_{OMLSA}(k,l) = \{G(k,l)\}^{p_{combinée}(k,l)} . G_{min}^{1-p_{combinée}(k,l)}$$

$$p_{combinée}(k,l) = p(k,l).p_{glob}(k)$$

[0081] En d'autres termes :

- en présence de parole au niveau de la trame, c'est-à-dire si $P_{speech}(k) > P_{seuil}$, l'expression conventionnelle du

calcul du gain OM-LSA reste inchangée ;

- en l'absence de parole au niveau de la trame, c'est-à-dire si $P_{speech}(k) < P_{seuil}$, les probabilités fréquentielles $p(k, l)$ seront en revanche pondérées par la probabilité globale $p_{glob}(k)$ faible, ce qui aura pour impact d'uniformiser les probabilités en diminuant leurs valeurs ;
- dans le cas asymptotique particulier $P_{speech}(k) = 0$, toutes les probabilités seront nulles et le débruitage sera totalement uniforme.

[0082] L'évaluation de la donnée globale $p_{glob}(k)$ est schématisée sur la Figure 2 par le bloc 32, qui reçoit en entrée les données P_{seuil} (valeur de seuil paramétrable) et $P_{speech}(k,l)$ (valeur elle-même calculée par le bloc 28, comme décrit plus haut), et délivre en sortie la valeur $p_{glob}(k)$ qui est appliquée à l'entrée ④ du bloc 24.

[0083] Ici encore, on utilise une donnée globale calculée au niveau de la trame pour affiner le calcul du gain fréquentiel de débruitage, et ceci en fonction du cas de figure rencontré (absence/présence de parole). En particulier, la donnée globale permet d'estimer la situation réelle rencontrée à l'échelle de la trame (phase de parole vs. phase de bruit seul), ce que la seule donnée fréquentielle ne permettrait pas de formuler. Et en présence de bruit seul, on peut se ramener à un débruitage uniforme, solution idéale car le bruit résiduel perçu ne sera alors jamais musical.

Résultats obtenus par l'algorithme de l'invention

[0084] Comme on vient de l'exposer, l'invention repose sur la mise en évidence de ce que le compromis débruitage/dégradation du signal repose sur un calcul de gain spectral (fonction d'un paramètre scalaire de gain minimal et d'une probabilité de présence de parole) dont le modèle est sous-optimal, et propose une formule impliquant une modulation temporelle de ces éléments de calcul du gain spectral, qui deviennent fonction de descripteurs temporels pertinents du signal de parole bruitée.

[0085] L'invention repose sur l'exploitation d'une donnée globale pour traiter de manière plus pertinente et adaptée chaque bande de fréquence, la dureté de débruitage étant rendue variable en fonction de la présence de parole sur une trame (on débruite plus quand le risque d'avoir une contrepartie est faible).

[0086] Dans l'algorithme OM-LSA conventionnel, chaque bande de fréquence est traitée de manière indépendante, et pour une fréquence donnée on n'intègre pas la connaissance *a priori* des autres bandes. Or, une analyse plus large qui observe l'ensemble de la trame pour calculer un indicateur global caractéristique de la trame (ici, un indicateur de présence de parole capable de discriminer même grossièrement phase de bruit seul et phase de parole) est un moyen utile et efficace pour affiner le traitement à l'échelle de la bande de fréquences.

[0087] Concrètement, dans un algorithme OM-LSA conventionnel, le gain de débruitage est généralement ajusté à une valeur de compromis, typiquement de l'ordre de 14 dB.

[0088] La mise en oeuvre de l'invention permet d'ajuster ce gain dynamiquement à une valeur variant entre 8 dB (en présence de parole) et 17 dB (en présence de bruit seul). La réduction de bruit est ainsi beaucoup plus énergique, et rend le bruit pratiquement imperceptible (et en tout état de cause non musical) en l'absence de parole dans la majeure partie des situations couramment rencontrées. Et même en présence de parole, le débruitage ne modifie pas la tonalité de la voix, dont le rendu reste naturel.

Revendications

1. Un procédé de débruitage d'un signal audio par application d'un algorithme à gain spectral variable fonction d'une probabilité de présence de parole, comportant les étapes successives suivantes :

   a) génération (10) de trames temporelles successives ($y(k)$) du signal audio bruité numérisé ($y(n)$) ;
   b) application d'une transformation de Fourier (12) aux trames générées à l'étape a), de manière à produire pour chaque trame temporelle de signal un spectre de signal ($Y(k,l)$) avec une pluralité de bandes de fréquences prédéterminées ;
   c) dans le domaine fréquentiel :

      c1) estimation (18), pour chaque bande de fréquences de chaque trame temporelle courante, d'une probabilité de présence de parole ($p(k,l)$) ;
      c3) calcul (16) d'un gain spectral ($G_{OMLSA}(k,l)$), propre à chaque bande de fréquence de chaque trame temporelle courante, en fonction de : i) une estimation de l'énergie du bruit dans chaque bande de fréquences, ii) la probabilité de présence de parole estimée à l'étape c1), et iii) une valeur scalaire de gain minimal (G_{min}) représentative d'un paramètre de dureté du débruitage ;
      c4) réduction sélective de bruit (14) par application à chaque bande de fréquences du gain calculé à l'étape

c3) ;

d) application d'une transformation de Fourier inverse (20) au spectre de signal ($\hat{X}(k,l)$) constitué des bandes de fréquences produites à l'étape c4), de manière à délivrer pour chaque spectre une trame temporelle de signal débruité ; et
e) reconstitution (22) d'un signal audio débruité à partir des trames temporelles délivrées à l'étape d),

procédé **caractérisé en ce que** :

- ladite valeur scalaire de gain minimal (G_{min}) est une valeur ($G_{min}(k)$) modulable de manière dynamique à chaque trame temporelle ($y(k)$) successive ; et
- le procédé comporte en outre, préalablement à l'étape c3) de calcul du gain spectral, une étape de :

c2) calcul (24), pour la trame temporelle courante ($y(k)$), de ladite valeur modulable ($G_{min}(k)$) en fonction d'une variable globale ($SNR_y\ (k)$; $P_{speech}(k)$; $VAD\ (k)$) observée au niveau de la trame temporelle courante pour toutes les bandes de fréquences ; et

- ledit calcul de l'étape c2) comprend l'application, pour la trame temporelle courante, d'un incrément/décrément ($\Delta G_{min}\ (k)$; $\Delta_1 G_{min}$, $\Delta_2 G_{min}$; ΔG_{min}) apporté à une valeur paramétrée nominale (G_{min}) dudit gain minimal.

2. Le procédé de la revendication 1, dans lequel ladite variable globale est un rapport signal sur bruit ($SNR_y\ (k)$) de la trame temporelle courante, évalué (26) dans le domaine temporel.

3. Le procédé de la revendication 2, dans lequel la valeur scalaire de gain minimal est calculée à l'étape c2) par application de la relation :

$$G_{min}(k) = G_{min} + \Delta G_{min}(SNR_y(k))$$

k étant l'indice de la trame temporelle courante,
$G_{min}(k)$ étant le gain minimal à appliquer à la trame temporelle courante,
G_{min} étant ladite valeur nominale paramétrée du gain minimal,
$\Delta G_{min}\ (k)$ étant ledit incrément/décrément apporté à G_{min}, et
$SNR_y\ (k)$ étant le rapport signal sur bruit de la trame temporelle courante.

4. Le procédé de la revendication 1, dans lequel ladite variable globale est une probabilité moyenne de parole ($P_{speech}(k)$), évaluée (28) au niveau de la trame temporelle courante.

5. Le procédé de la revendication 4, dans lequel la valeur scalaire de gain minimal est calculée à l'étape c2) par application de la relation :

$$G_{min}(k) = G_{min} + \left(P_{speech}(k) - 1\right).\Delta_1 G_{min} + P_{speech}(k).\Delta_2 G_{min}$$

k étant l'indice de la trame temporelle courante,
$G_{min}(k)$ étant le gain minimal à appliquer à la trame temporelle courante,
G_{min} étant ladite valeur nominale paramétrée du gain minimal,
$P_{speech}(k)$ étant la probabilité moyenne de parole évaluée au niveau de la trame temporelle courante,
$\Delta_1 G_{min}$ étant ledit incrément/décrément, apporté à G_{min} en phase de bruit, et
$\Delta_2 G_{min}$ étant ledit incrément/décrément, apporté à G_{min} en phase de parole.

6. Le procédé de la revendication 4, dans lequel la probabilité moyenne de parole est évaluée au niveau de la trame temporelle courante par application de la relation :

$$P_{speech}(k) = \frac{1}{N}\sum_{l}^{N} p(k,l)$$

l étant l'indice de la bande de fréquences,
N étant le nombre de bandes de fréquences dans le spectre, et
p(*k*,*l*) étant la probabilité de présence de parole de la bande de fréquences d'indice *l* de la trame temporelle courante.

7. Le procédé de la revendication 1, dans lequel ladite variable globale est un signal booléen de détection d'activité vocale (*VAD* (*k*)) pour la trame temporelle courante, évalué (30) dans le domaine temporel par analyse de la trame temporelle et/ou au moyen d'un détecteur externe.

8. Le procédé de la revendication 7, dans lequel la valeur scalaire de gain minimal est calculée à l'étape c2) par application de la relation :

$$G_{min}(k) = G_{min} + VAD(k).\Delta G_{min}$$

k étant l'indice de la trame temporelle courante,
$G_{min}(k)$ étant le gain minimal à appliquer à la trame temporelle courante, G_{min} étant ladite valeur nominale paramétrée du gain minimal,
VAD (*k*) étant la valeur du signal booléen de détection d'activité vocale pour la trame temporelle courante, et
ΔG_{min} étant ledit incrément/décrément apporté à G_{min}.

Patentansprüche

1. Verfahren zur Rauschunterdrückung in einem Audiosignal durch Anwendung eines Algorithmus mit variabler Spektralverstärkung als Funktion eines wahrscheinlichen Vorhandenseins von Sprache, umfassend die folgenden aufeinanderfolgenden Schritte:

a) Erzeugung (10) von aufeinanderfolgenden Zeitrahmen *(y(k))* des digitalen verrauschten Audiosignals *(y(n))*;
b) Anwendung einer Fouriertransformation (12) auf die Rahmen, die in Schritt a) erzeugt wurden, dergestalt, dass für jeden Zeitrahmen des Signals ein Signalspektrum *(Y(k,l))* mit einer Vielzahl von zuvor bestimmten Frequenzbändern erzeugt wird;
c) in der Frequenzdomäne:

c1) Schätzung (18) für jedes Frequenzband jedes laufenden Zeitrahmens eines wahrscheinlichen Vorhandenseins von Sprache *(p(k,l));*
c3) Berechnung (16) einer Spektralverstärkung *($G_{OMLSA}(k,l)$),* die für jedes Frequenzband von jedem laufenden Zeitrahmen typisch ist, in Abhängigkeit von: i) einer Schätzung der Energie des Rauschens in jedem Frequenzband, ii) dem wahrscheinlichen Vorhandensein von Sprache, das in Schritt c1) geschätzt wurde, und iii) einem Skalarwert minimaler Verstärkung (G_{min}), der ein Parameter der Härte der Rauschunterdrückung repräsentiert;
c4) selektive Rauschreduzierung (14) durch Anwendung der Verstärkung, die in Schritt c3) berechnet wurde, auf jedes Frequenzband;

d) Anwendung einer umgekehrten Fouriertransformation (20) auf das Signalspektrum *($\hat{X}(k,1)$),* das aus den Frequenzbändern besteht, die in Schritt c4) erzeugt wurden, dergestalt, dass für jedes Spektrum ein Zeitrahmen des rauschunterdrückten Signals geliefert wird; und
e) Wiedergewinnung (22) eines rauschunterdrückten Audiosignals ausgehend von den Zeitrahmen, die in Schritt d) geliefert wurden,

wobei das Verfahren **dadurch gekennzeichnet ist, dass**:

- der Skalarwert der minimalen Verstärkung *(G_{min})* ein Wert *($G_{min}(k)$)* ist, der auf dynamische Art in jedem aufeinanderfolgenden Zeitrahmen *(y(k))* modulierbar ist; und
- das Verfahren außerdem, vor dem Schritt c3) der Berechnung der Spektralverstärkung, einen Schritt zur:

c2) Berechnung (24) für den laufenden Zeitrahmen (*y*(k)) des modulierbaren Wertes (G_{min}(k)) in Abhängigkeit einer globalen Variablen *($SNR_y(k)$; $P_{speech}(k)$; VAD (k))* aufweist, die in dem laufenden Zeitrahmen für alle Frequenzbänder beobachtet wird; und

- die Berechnung des Schritts c2) die Anwendung für den laufenden Zeitrahmen eines Inkrements/Dekrements *($\Delta G_{min}(k)$; $\Delta_1 G_{min}$, $\Delta_2 G_{min}$; AGmin)* aufweist, das auf einen nominalen Parameterwert (G_{min}) während der minimalen Verstärkung angewendet wird.

2. Verfahren nach Anspruch 1, wobei die globale Variable ein Signal-Rausch-Verhältnis *($SNR_y(k)$)* des laufenden Zeitrahmens ist, welches in der Zeitdomäne bewertet (26) wird.

3. Verfahren nach Anspruch 2, wobei der Skalarwert der minimalen Verstärkung in Schritt c2) durch Anwendung des Verhältnisses berechnet wird:

$$G_{min}(k) = G_{min} + \Delta G_{min}(SNR_y(k))$$

wobei k der Index des laufenden Zeitrahmens ist,
wobei $G_{min}(k)$ die minimale Verstärkung ist, die auf den laufenden Zeitrahmen anzuwenden ist,
wobei G_{min} der nominale Parameterwert der minimalen Verstärkung ist,
wobei $\Delta_{Gmin}(k)$ das Inkrement/Dekrement ist, das auf Gmin angewendet wird, und
wobei $SNR_y(k)$ das Signal-Rausch-Verhältnis des laufenden Zeitrahmens ist.

4. Verfahren nach Anspruch 1, wobei die globale Variable eine durchschnittliche Wahrscheinlichkeit von Sprache *($P_{speech}(k)$)* ist, die in dem laufenden Zeitrahmen bewertet (28) wird.

5. Verfahren nach Anspruch 4, wobei der Skalarwert der minimalen Verstärkung in Schritt c2) durch Anwendung des Verhältnisses berechnet wird:

$$G_{min}(k) = G_{min} + (P_{speech}(k) - 1) \cdot \Delta_1 G_{min} + P_{speech}(k) \cdot \Delta_2 G_{min}$$

wobei *k* der Index des laufenden Zeitrahmens ist,
wobei *Gmin (k)* die minimale Verstärkung ist, die auf den laufenden Zeitrahmen anzuwenden ist,
wobei G_{min} der nominale Parameterwert der minimalen Verstärkung ist,
wobei $P_{speech}(k)$ die durchschnittliche Wahrscheinlichkeit von Sprache ist, die in dem laufenden Zeitrahmen bewertet wird,
wobei $\Delta_1 G_{min}$ das Inkrement/Dekrement ist, das auf G_{min} in der Rauschphase angewendet wird, und
wobei $\Delta_2 G_{min}$ das Inkrement/Dekrement ist, das auf *Gmin* in der Sprachphase angewendet wird.

6. Verfahren nach Anspruch 4, wobei die durchschnittliche Wahrscheinlichkeit von Sprache in dem laufenden Zeitrahmen durch Anwendung des Verhältnisses bewertet wird:

$$P_{speech}(k) = \frac{1}{N} \sum_{1}^{N} p(k,l)$$

wobei 1 der Index des Frequenzbandes ist,
wobei N die Anzahl an Frequenzbändern in dem Spektrum ist, und
wobei *p(k,l))* die Wahrscheinlichkeit des Vorhandenseins von Sprache des Frequenzbandes mit Index 1 des laufenden Zeitrahmens ist.

7. Verfahren nach Anspruch 1, wobei die globale Variable ein boolesches Signal zur Erkennung von Sprachaktivität *(VAD(k))* für den laufenden Zeitrahmen ist, welches in der Zeitdomäne durch Analyse des Zeitrahmens und/oder mithilfe einer externen Erkennung bewertet (30) wird.

8. Verfahren nach Anspruch 7, wobei der Skalarwert der minimalen Verstärkung in Schritt c2) durch Anwendung des Verhältnisses berechnet wird:

$$G_{min}(k) = G_{min} + VAD(k) \cdot \Delta G_{min}$$

wobei k der Index des laufenden Zeitrahmens ist, wobei $G_{min}(k)$ die minimale Verstärkung ist, die auf den laufenden Zeitrahmen anzuwenden ist,
wobei G_{min} der nominale Parameterwert der minimalen Verstärkung ist,
wobei *VAD(k)* der Wert des booleschen Erkennungssignals von Sprachaktivität für den laufenden Zeitrahmen ist, und
wobei ΔG_{min} das Inkrement/Dekrement ist, das auf *Gmin* angewendet wird.

Claims

1. A method for denoising an audio signal by applying an algorithm with a variable spectral gain, which is function of a speech presence probability, including the following successive steps:

a) generating (10) successive time frames *(y(k))* of the digitized noisy audio signal *(y(n))* ;
b) applying a Fourier transform (12) to the frames generated at step a), so as to produce for each signal time frame a signal spectrum (Y(k, l)) with a plurality of predetermined frequency bands;
c) in the frequency domain:

c1) estimating (18), for each frequency band of each current time frame, a speech presence probability (p(k, l)) ;
c3) calculating (16) a spectral gain $(G_{omLsA}(k,l))$, specific to each frequency band of each current time frame, as a function of: i) an estimation of the noise energy in each frequency band, ii) the speech presence probability estimated at step c1), and iii) a scalar value of minimum gain (G_{min}) representative of a parameter of hardness of the denoising;
c4) performing (14) a selective noise reduction by applying to each frequency band of the gain calculated at step c3);

d) applying an inverse Fourier transform (20) to the signal spectrum ($\hat{X}(k,l)$) consisted of the frequency bands produced at step c4), so as to deliver for each spectrum a time frame of denoised signal; and
e) reconstructing (22) a denoised audio signal from the time frames delivered at step d),

the method being **characterized in that**:

- said scalar value of minimum gain (G_{min}) is a value $(G_{min}(k))$ that is dynamically modulatable at each successive time frame *(y(k))* ; and
- the method further includes, previously to step c3) of calculating the spectral gain, a step of:

c2) calculating (24), for the current time frame *(y(k))*, said modulatable value $(G_{min}(k))$, as a function of a global variable $(SNR_y(k))$; $(P_{speech}(k))$; *(VAD(k))* observed at the current time frame for all the frequency bands; and

- said calculation of step c2) comprises apply, for the current time frame, an increment/decrement $(\Delta G_{min}(k);$ $\Delta_1 G_{min}, \Delta_2 G_{min}; \Delta G_{min})$ added to a parameterized nominal value (G_{min}) of said minimum gain.

2. The method of claim 1, wherein said global variable is a signal-to-noise ratio $(SNR_y(k))$ of the current time frame, evaluated (26) in the time domain.

3. The method of claim 2, wherein the scalar value of minimum gain is calculated at step c2) by application of the relation:

$$G_{min}(k) = G_{min} + \Delta G_{min}(SNR_y(k))$$

k being the index of the current time frame,
$G_{min}(k)$ being the minimum gain to be applied to the current time frame,
G_{min} being said parameterized nominal value of the minimum gain,
$\Delta G_{min}(k)$ being said increment/decrement added to G_{min}, and
$SNR_y(k)$ being the signal-to-noise ratio of the current time frame.

4. The method of claim 1, wherein said global variable is an average speech probability ($P_{speech}(k)$), evaluated (28) at the current time frame.

5. The method of claim 4, wherein the scalar value of minimum gain is calculated at step c2) by application of the relation:

$$G_{min}(k) = G_{min} + (P_{speech}(k) - 1) \cdot \Delta_1 G_{min} + P_{speech}(k) \cdot \Delta_2 G_{min}$$

k being the index of the current time frame,
$G_{min}(k)$ being the minimum gain to be applied to the current time frame,
G_{min} being said parameterized nominal value of the minimum gain,
$P_{speech}(k)$ being the average speech probability evaluated at the current time frame,
$\Delta_1 G_{min}$ being said increment/decrement added to G_{min} in phase of noise, and
$\Delta_2 G_{min}$ being said increment/decrement added to G_{min} in phase of speech.

6. The method of claim 4, wherein the average speech probability is evaluated at the current time frame by application of the relation:

$$P_{speech}(k) = \frac{1}{N} \sum_{l}^{N} p(k,l)$$

l being the index of the frequency band,
N being the number of frequency bands in the spectrum, and $p(k,l)$ being the speech presence probability of the frequency band of index l of the current time frame.

7. The method of claim 1, wherein said global variable is a Boolean signal of voice activity detection $(VAD(k))$ for the current time frame, evaluated (30) in the time domain by analysis of the time frame and/or by means of an external detector.

8. The method of claim 7, wherein the scalar value of minimum gain is calculated at step c2) by application of the relation:

$$G_{min}(k) = G_{min} + VAD(k) \cdot \Delta G_{min}$$

k being the index of the current time frame,
$G_{min}(k)$ being the minimum gain to be applied to the current time frame,

G_{min} being said parameterized nominal value of the minimum gain,
VAD(k) being the value of the Boolean signal of voice activity detection for the current time frame, and
ΔG_{min} being said increment/decrement added to G_{min}.

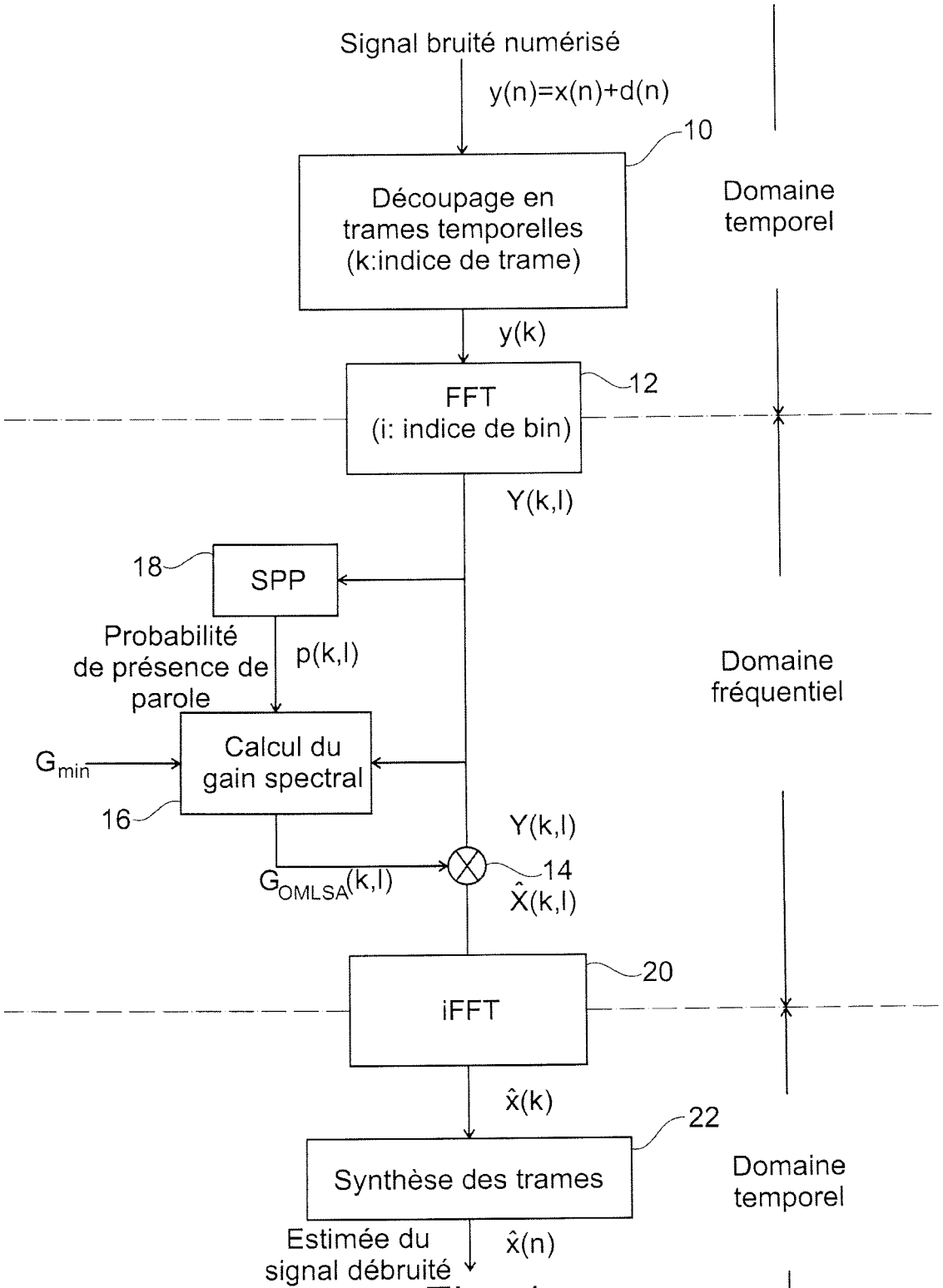

Fig. 1
(Technique antérieure)

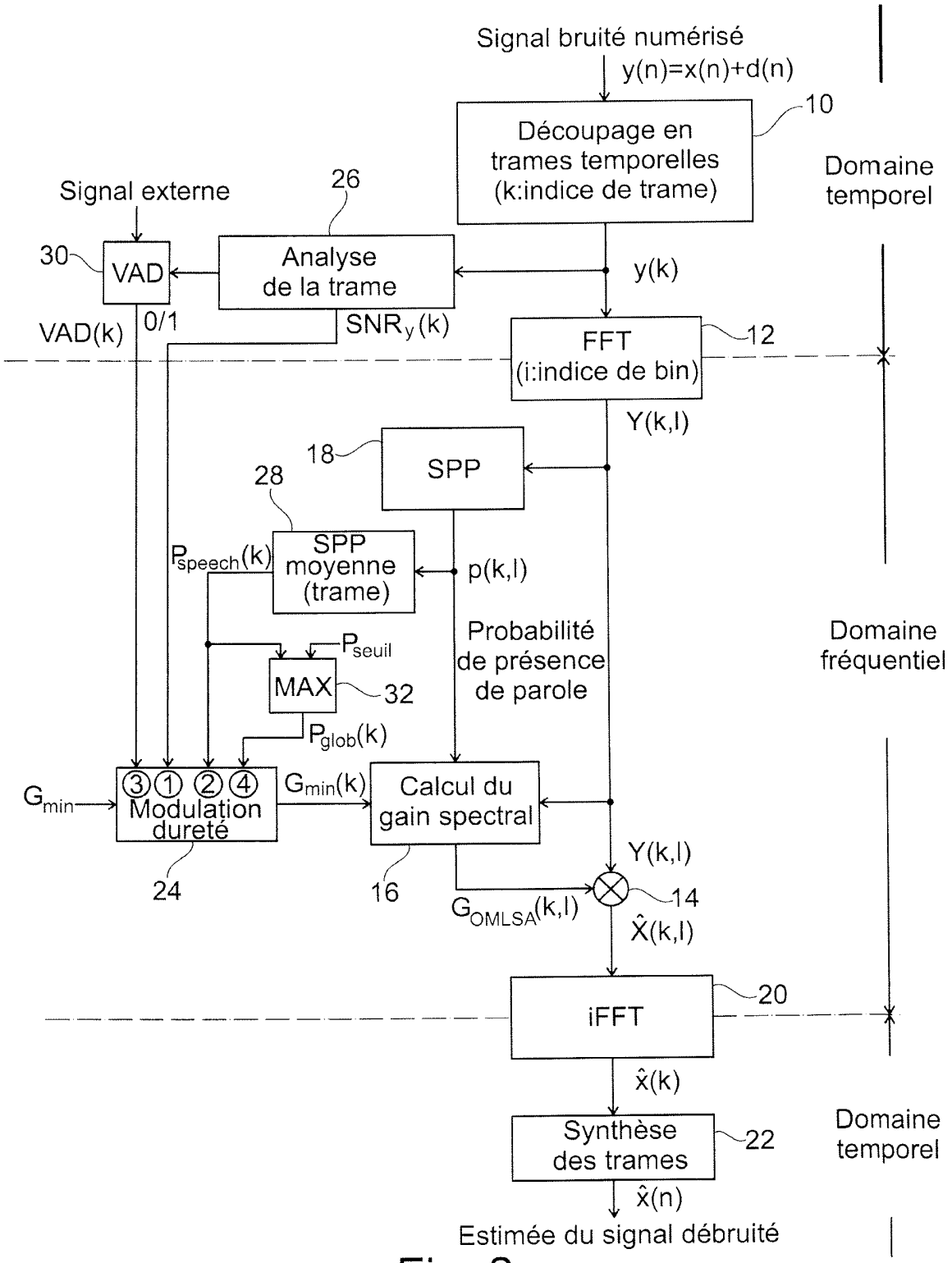
Signal bruité numérisé
y(n)=x(n)+d(n)
Découpage en trames temporelles (k:indice de trame)
10
Domaine temporel
Signal externe
26
30
VAD
Analyse de la trame
y(k)
VAD(k)
0/1
SNR$_y$(k)
FFT (i:indice de bin)
12
Y(k,l)
18
SPP
28
SPP moyenne (trame)
P$_{speech}$(k)
p(k,l)
Probabilité de présence de parole
Domaine fréquentiel
P$_{seuil}$
MAX
32
P$_{glob}$(k)
G$_{min}$
3 1 2 4
Modulation dureté
G$_{min}$(k)
Calcul du gain spectral
24
16
G$_{OMLSA}$(k,l)
Y(k,l)
14
$\hat{X}$(k,l)
iFFT
20
$\hat{x}$(k)
Synthèse des trames
22
Domaine temporel
$\hat{x}$(n)
Estimée du signal débruité

Fig. 2

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 7454010 B1 **[0014]**
- WO 2007099222 A1, Parrot **[0015] [0050]**
- EP 2530672 A1 **[0071]**
- EP 2518724 A1 **[0072]**

Littérature non-brevet citée dans la description

- **I. COHEN ; B. BERDUGO.** Speech Enhancement for Non-Stationary Noise Environments. *Signal Processing,* Novembre 2001, vol. 81 (11), 2403-2418 **[0013]**
- **I. COHEN.** Optimal Speech Enhancement Under Signal Presence Uncertainty Using Log-Spectral Amplitude Estimator. *IEEE Signal Processing Letters,* Avril 2002, vol. 9 (4), 113-116 **[0013] [0044]**
- **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, Hong-Kong,* Avril 2003, 233-236 **[0048]**