

(19)



Europäisches Patentamt
European Patent Office
Office européen des brevets



(11)

EP 1 016 071 B1

(12)

FASCICULE DE BREVET EUROPEEN

(45) Date de publication et mention
de la délivrance du brevet:

16.01.2002 Bulletin 2002/03

(21) Numéro de dépôt: **98943998.9**

(22) Date de dépôt: **16.09.1998**

(51) Int Cl.7: **G10L 11/02, G10L 11/06**

(86) Numéro de dépôt international:
PCT/FR98/01979

(87) Numéro de publication internationale:
WO 99/14737 (25.03.1999 Gazette 1999/12)

(54) **PROCEDE ET DISPOSITIF DE DETECTION D'ACTIVITE VOCALE**

VERFAHREN UND VORRICHTUNG ZUR SPRACHDETEKTION

METHOD AND APPARATUS FOR DETECTING SPEECH ACTIVITY

(84) Etats contractants désignés:
CH DE FI GB LI SE

(30) Priorité: **18.09.1997 FR 9711640**

(43) Date de publication de la demande:
05.07.2000 Bulletin 2000/27

(73) Titulaires:

- **Matra Nortel Communications**
29100 Quimper (FR)
Etats contractants désignés:
DE FI GB SE
- **Eads Defence and Security Networks**
78180 Montigny Le Bretonneux (FR)
Etats contractants désignés:
CH LI

(72) Inventeurs:

- **LOCKWOOD, Philip**
F-95490 Vaureal (FR)
- **LUBIARZ, Stéphane**
F-95520 Osny (FR)

(74) Mandataire: **Loisel, Bertrand et al**
Cabinet Plasseraud, 84, rue d'Amsterdam
75440 Paris Cédex 09 (FR)

(56) Documents cités:

DE-A- 4 012 349 **US-A- 5 659 622**

- **PATENT ABSTRACTS OF JAPAN vol. 095, no.**
006, 31 juillet 1995 & JP 07 074709 A (SONY), 17
mars 1995 -& US 5 732 390 A (KATAYANAGI ET
AL.) 24 mars 1998

EP 1 016 071 B1

Il est rappelé que: Dans un délai de neuf mois à compter de la date de publication de la mention de la délivrance du brevet européen, toute personne peut faire opposition au brevet européen délivré, auprès de l'Office européen des brevets. L'opposition doit être formée par écrit et motivée. Elle n'est réputée formée qu'après paiement de la taxe d'opposition. (Art. 99(1) Convention sur le brevet européen).

Description

[0001] La présente invention concerne les techniques numériques de traitement de signaux de parole. Elle concerne plus particulièrement les techniques faisant appel à une détection d'activité vocale afin d'effectuer des traitements différenciés selon que le signal supporte ou non une activité vocale.

[0002] Les techniques numériques en question relèvent de domaines variés : codage de la parole pour la transmission ou le stockage, reconnaissance de la parole, diminution du bruit, annulation d'écho...

[0003] Les méthodes de détection d'activité vocale ont pour principale difficulté la distinction entre l'activité vocale et le bruit qui l'accompagne. Le recours à une technique de débruitage classique ne permet pas de traiter cette difficulté, puisque ces techniques font elles-mêmes appel à des estimations du bruit qui dépendent du degré d'activité vocale du signal. Ce problème a par exemple été décrit dans le document US-A-5659622.

[0004] Un but principal de la présente invention est d'améliorer la robustesse au bruit des méthodes de détection d'activité vocale. Pour atteindre ce but, un procédé comme indiqué dans la revendication 1 est proposé.

[0005] L'invention propose ainsi un procédé de détection d'activité vocale dans un signal de parole numérique traité par trames successives, dans lequel on soumet le signal de parole à un débruitage en tenant compte d'estimations du bruit compris dans le signal, mises à jour pour chaque trame d'une manière dépendante d'au moins un degré d'activité vocale déterminé pour ladite trame. Selon l'invention, on procède à un débruitage a priori du signal de parole de chaque trame sur la base d'estimations du bruit obtenues lors du traitement d'au moins une trame précédente, et on analyse les variations d'énergie du signal débruité a priori pour détecter le degré d'activité vocale de ladite trame.

[0006] Le fait de procéder à la détection d'activité vocale (selon une méthode qui peut généralement être toute méthode connue) sur la base d'un signal débruité a priori améliore sensiblement les performances de cette détection lorsque le bruit environnant est relativement important.

[0007] Dans la suite de la présente description, on illustrera le procédé de détection d'activité vocale selon l'invention dans un système de débruitage d'un signal de parole. On comprendra que ce procédé peut trouver des applications dans de nombreux autres types de traitement numérique de la parole dans lesquels on souhaite disposer d'une information sur le degré d'activité vocale du signal traité : codage, reconnaissance, annulation d'écho...

[0008] D'autres particularités et avantages de la présente invention apparaîtront dans la description ci-après d'exemples de réalisation non limitatifs, en référence aux dessins annexés, dans lesquels :

- la figure 1 est un schéma synoptique d'un système de débruitage mettant en oeuvre la présente invention ;
- les figures 2 et 3 sont des organigrammes de procédures utilisées par un détecteur d'activité vocale du système de la figure 1 ;
- la figure 4 est un diagramme représentant les états d'un automate de détection d'activité vocale ;
- la figure 5 est un graphique illustrant les variations d'un degré d'activité vocale ;
- la figure 6 est un schéma synoptique d'un module de surestimation du bruit du système de la figure 1 ;
- la figure 7 est un graphique illustrant le calcul d'une courbe de masquage ; et
- la figure 8 est un graphique illustrant l'exploitation des courbes de masquage dans le système de la figure 1.

[0009] Le système de débruitage représenté sur la figure 1 traite un signal numérique de parole s . Un module de fenêtrage 10 met ce signal s sous forme de fenêtres ou trames successives, constituées chacune d'un nombre N d'échantillons de signal numérique. De façon classique, ces trames peuvent présenter des recouvrements mutuels. Dans la suite de la présente description, on considérera, sans que ceci soit limitatif, que les trames sont constituées de $N=256$ échantillons à une fréquence d'échantillonnage F_e de 8 kHz, avec une pondération de Hamming dans chaque fenêtre, et des recouvrements de 50% entre fenêtres consécutives.

[0010] La trame de signal est transformée dans le domaine fréquentiel par un module 11 appliquant un algorithme classique de transformée de Fourier rapide (TFR) pour calculer le module du spectre du signal. Le module 11 délivre alors un ensemble de $N=256$ composantes fréquentielles du signal de parole, notées $S_{n,f}$, où n désigne le numéro de la trame courante, et f une fréquence du spectre discret. Du fait des propriétés des signaux numériques dans le domaine fréquentiel, seuls les $N/2=128$ premiers échantillons sont utilisés.

[0011] Pour calculer les estimations du bruit contenu dans le signal s , on n'utilise pas la résolution fréquentielle disponible en sortie de la transformée de Fourier rapide, mais une résolution plus faible, déterminée par un nombre I de bandes de fréquences couvrant la bande $[0, F_e/2]$ du signal. Chaque bande i ($1 \leq i \leq I$) s'étend entre une fréquence inférieure $f(i-1)$ et une fréquence supérieure $f(i)$, avec $f(0)=0$, et $f(I)=F_e/2$. Ce découpage en bandes de fréquences peut être uniforme ($f(i)-f(i-1)=F_e/2I$). Il peut également être non uniforme (par exemple selon une échelle de barks). Un module 12 calcule les moyennes respectives des composantes spectrales $S_{n,f}$ du signal de parole par bandes, par exemple par une pondération uniforme telle que :

$$S_{n,i} = \frac{1}{f(i) - f(i-1)} \sum_{f \in [f(i-1), f(i)]} S_{n,f} \quad (1)$$

[0012] Ce moyennage diminue les fluctuations entre les bandes en moyennant les contributions du bruit dans ces bandes, ce qui diminuera la variance de l'estimateur de bruit. En outre, ce moyennage permet une forte diminution de la complexité du système.

[0013] Les composantes spectrales moyennées $S_{n,i}$ sont adressées à un module 15 de détection d'activité vocale et à un module 16 d'estimation du bruit. Ces deux modules 15, 16 fonctionnent conjointement, en ce sens que des degrés d'activité vocale $\gamma_{n,i}$ mesurés pour les différentes bandes par le module 15 sont utilisés par le module 16 pour estimer l'énergie à long terme du bruit dans les différentes bandes, tandis que ces estimations à long terme $\hat{E}_{n,i}$ sont utilisées par le module 15 pour procéder à un débruitage a priori du signal de parole dans les différentes bandes pour déterminer les degrés d'activité vocale $\gamma_{n,i}$.

[0014] Le fonctionnement des modules 15 et 16 peut correspondre aux organigrammes représentés sur les figures 2 et 3.

[0015] Aux étapes 17 à 20, le module 15 procède au débruitage a priori du signal de parole dans les différentes bandes i pour la trame de signal n . Ce débruitage a priori est effectué selon un processus classique de soustraction spectrale non linéaire à partir d'estimations du bruit obtenues lors d'une ou plusieurs trames précédentes. A l'étape 17, le module 15 calcule, avec la résolution des bandes i , la réponse en fréquence $H_{p_{n,i}}$ du filtre de débruitage a priori, selon la formule :

$$H_{p_{n,i}} = \frac{S_{n,i} - \alpha'_{n-\tau_1,i} \cdot \hat{B}_{n-\tau_1,i}}{S_{n-\tau_2,i}} \quad (2)$$

où τ_1 et τ_2 sont des retards exprimés en nombre de trames ($\tau_1 \geq 1$, $\tau_2 \geq 0$), et $\alpha'_{n,i}$ est un coefficient de surestimation du bruit dont la détermination sera expliquée plus loin. Le retard τ_1 peut être fixe (par exemple $\tau_1=1$) ou variable. Il est d'autant plus faible qu'on est confiant dans la détection d'activité vocale.

[0016] Aux étapes 18 à 20, les composantes spectrales $\hat{E}_{p_{n,i}}$ sont calculées selon :

$$\hat{E}_{p_{n,i}} = \max \left\{ H_{p_{n,i}} \cdot S_{n,i}, \beta p_i \cdot \hat{E}_{n-\tau_1,i} \right\} \quad (3)$$

où βp_i est un coefficient de plancher proche de 0, servant classiquement à éviter que le spectre du signal débruité prenne des valeurs négatives ou trop faibles qui provoqueraient un bruit musical.

[0017] Les étapes 17 à 20 consistent donc essentiellement à soustraire du spectre du signal une estimation, majorée par le coefficient $\alpha'_{n-\tau_1,i}$, du spectre du bruit estimé a priori.

[0018] A l'étape 21, le module 15 calcule l'énergie du signal débruité a priori dans les différentes bandes i pour la trame n : $E_{n,i} = \hat{E}_{p_{n,i}}^2$. Il calcule aussi une moyenne globale $E_{n,0}$ de l'énergie du signal débruité a priori, par une somme des énergies par bande $E_{n,i}$, pondérée par les largeurs de ces bandes. Dans les notations ci-dessous, l'indice $i=0$ sera utilisé pour désigner la bande globale du signal.

[0019] Aux étapes 22 et 23, le module 15 calcule, pour chaque bande i ($0 \leq i \leq I$), une grandeur $\Delta E_{n,i}$ représentant la variation à court terme de l'énergie du signal débruité dans la bande i , ainsi qu'une valeur à long terme $\bar{E}_{n,i}$ de l'énergie du signal débruité dans la bande i . La grandeur $\Delta E_{n,i}$ peut être calculée par une formule simplifiée de dérivation :

$$\Delta E_{n,i} = \left| \frac{E_{n-4,i} + E_{n-3,i} - E_{n-1,i} - E_{n,i}}{10} \right|$$

. Quant à l'énergie à long terme $E_{n,i}$, elle peut être calculée à l'aide d'un facteur d'oubli $B1$ tel que $0 < B1 < 1$, à savoir $\bar{E}_{n,i} = B1 \cdot \bar{E}_{n-1,i} + (1-B1) \cdot E_{n,i}$.

[0020] Après avoir calculé les énergies $E_{n,i}$ du signal débruité, ses variations à court terme $\Delta E_{n,i}$ et ses valeurs à long terme $\bar{E}_{n,i}$ de la manière indiquée sur la figure 2, le module 15 calcule, pour chaque bande i ($0 \leq i \leq I$), une valeur p_i représentative de l'évolution de l'énergie du signal débruité. Ce calcul est effectué aux étapes 25 à 36 de la figure 3, exécutées pour chaque bande i entre $i=0$ et $i=I$. Ce calcul fait appel à un estimateur à long terme de l'enveloppe

du bruit ba_i , à un estimateur interne bi_i et à un compteur de trames bruitées b_i .

[0021] A l'étape 25, la grandeur $\Delta E_{n,i}$ est comparée à un seuil ε_1 . Si le seuil ε_1 n'est pas atteint, le compteur b_i est incrémenté d'une unité à l'étape 26. A l'étape 27, l'estimateur à long terme ba_i est comparé à la valeur de l'énergie lissée $E_{n,i}$. Si $ba_i \geq \bar{E}_{n,i}$, l'estimateur ba_i est pris égal à la valeur lissée $\bar{E}_{n,i}$ à l'étape 28, et le compteur b_i est remis à zéro. La grandeur ρ_i , qui est prise égale au rapport $ba_i/\bar{E}_{n,i}$ (étape 36), est alors égale à 1.

[0022] Si l'étape 27 montre que $ba_i < \bar{E}_{n,i}$, le compteur b_i est comparé à une valeur limite b_{max} à l'étape 29. Si $b_i > b_{max}$, le signal est considéré comme trop stationnaire pour supporter de l'activité vocale. L'étape 28 précitée, qui revient à considérer que la trame ne comporte que du bruit, est alors exécutée. Si $b_i \leq b_{max}$ à l'étape 29, l'estimateur interne bi_i est calculé à l'étape 33 selon :

$$bi_i = (1 - Bm) \cdot \bar{E}_{n,i} + Bm \cdot ba_i \quad (4)$$

Dans cette formule, Bm représente un coefficient de mise à jour compris entre 0,90 et 1. Sa valeur diffère selon l'état d'un automate de détection d'activité vocale (étapes 30 à 32). Cet état δ_{n-1} est celui déterminé lors du traitement de la trame précédente. Si l'automate est dans un état de détection de parole ($\delta_{n-1}=2$ à l'étape 30), le coefficient Bm prend une valeur Bmp très proche de 1 pour que l'estimateur du bruit soit très faiblement mis à jour en présence de parole. Dans le cas contraire, le coefficient Bm prend une valeur Bms plus faible, pour permettre une mise à jour plus significative de l'estimateur de bruit en phase de silence. A l'étape 34, l'écart $ba_i - bi_i$ entre l'estimateur à long terme et l'estimateur interne du bruit est comparé à un seuil ε_2 . Si le seuil ε_2 n'est pas atteint, l'estimateur à long terme ba_i est mis à jour avec la valeur de l'estimateur interne bi_i à l'étape 35. Sinon, l'estimateur à long terme ba_i reste inchangé. On évite ainsi que de brutales variations dues à un signal de parole conduisent à une mise à jour de l'estimateur de bruit.

[0023] Après avoir obtenu les grandeurs ρ_i , le module 15 procède aux décisions d'activité vocale à l'étape 37. Le module 15 met d'abord à jour l'état de l'automate de détection selon la grandeur ρ_0 calculée pour l'ensemble de la bande du signal. Le nouvel état δ_n de l'automate dépend de l'état précédent δ_{n-1} et de ρ_0 , de la manière représentée sur la figure 4.

[0024] Quatre états sont possibles : $\delta=0$ détecte le silence, ou absence de parole ; $\delta=2$ détecte la présence d'une activité vocale ; et les états $\delta=1$ et $\delta=3$ sont des états intermédiaires de montée et de descente. Lorsque l'automate est dans l'état de silence ($\delta_{n-1}=0$), il y reste si ρ_0 ne dépasse pas un premier seuil SE1, et il passe dans l'état de montée dans le cas contraire. Dans l'état de montée ($\delta_{n-1}=1$), il revient dans l'état de silence si ρ_0 est plus petit que le seuil SE1, il passe dans l'état de parole si ρ_0 est plus grand qu'un second seuil SE2 plus grand que le seuil SE1, et il reste dans l'état de montée si $SE1 \leq \rho_0 \leq SE2$. Lorsque l'automate est dans l'état de parole ($\delta_{n-1}=2$), il y reste si ρ_0 dépasse un troisième seuil SE3 plus petit que le seuil SE2, et il passe dans l'état de descente dans le cas contraire. Dans l'état de descente ($\delta_{n-1}=3$), l'automate revient dans l'état de parole si ρ_0 est plus grand que le seuil SE2, il revient dans l'état de silence si ρ_0 est en deçà d'un quatrième seuil SE4 plus petit que le seuil SE2, et il reste dans l'état de descente si $SE4 \leq \rho_0 \leq SE2$.

[0025] A l'étape 37, le module 15 calcule également les degrés d'activité vocale $\gamma_{n,i}$ dans chaque bande $i \geq 1$. Ce degré $\gamma_{n,i}$ est de préférence un paramètre non binaire, c'est-à-dire que la fonction $\gamma_{n,i} = g(\rho_i)$ est une fonction variant continûment entre 0 et 1 en fonction des valeurs prises par la grandeur ρ_i . Cette fonction a par exemple l'allure représentée sur la figure 5.

[0026] Le module 16 calcule les estimations du bruit par bande, qui seront utilisées dans le processus de débruitage, en utilisant les valeurs successives des composantes $S_{n,i}$ et des degrés d'activité vocale $\gamma_{n,i}$. Ceci correspond aux étapes 40 à 42 de la figure 3. A l'étape 40, on détermine si l'automate de détection d'activité vocale vient de passer de l'état de montée à l'état de parole. Dans l'affirmative, les deux dernières estimations $B_{n-1,i}$ et $B_{n-2,i}$ précédemment calculées pour chaque bande $i \geq 1$ sont corrigées conformément à la valeur de l'estimation précédente $B_{n-3,i}$. Cette correction est effectuée pour tenir compte du fait que, dans la phase de montée ($\delta=1$), les estimations à long terme de l'énergie du bruit dans le processus de détection d'activité vocale (étapes 30 à 33) ont pu être calculées comme si le signal ne comportait que du bruit ($Bm=Bms$), de sorte qu'elles risquent d'être entachées d'erreur.

[0027] A l'étape 42, le module 16 met à jour les estimations du bruit par bande selon les formules :

$$\tilde{S}_{n,i} = \lambda_B \cdot \hat{S}_{n-1,i} + (1 - \lambda_B) \cdot S_{n,i} \quad (5)$$

$$\hat{S}_{n,i} = \gamma_{n,i} \cdot \hat{S}_{n-1,i} + (1 - \gamma_{n,i}) \cdot \tilde{S}_{n,i} \quad (6)$$

où λ_B désigne un facteur d'oubli tel que $0 < \lambda_B < 1$. La formule (6) met en évidence la prise en compte du degré d'activité vocale non binaire $\gamma_{n,1}$.

[0028] Comme indiqué précédemment, les estimations à long terme du bruit $\hat{B}_{n,i}$ font l'objet d'une surestimation, par un module 45 (figure 1), avant de procéder au débruitage par soustraction spectrale non linéaire. Le module 45 calcule le coefficient de surestimation $\alpha'_{n,i}$ précédemment évoqué, ainsi qu'une estimation majorée $B'_{n,i}$ qui correspond essentiellement à $\alpha'_{n,i} \cdot \hat{B}_{n,i}$.

[0029] L'organisation du module de surestimation 45 est représentée sur la figure 6. L'estimation majorée $B'_{n,i}$ est obtenue en combinant l'estimation à long terme $\hat{B}_{n,i}$ et une mesure $\Delta B_{n,i}^{\max}$ de la variabilité de la composante du bruit dans la bande i autour de son estimation à long terme. Dans l'exemple considéré, cette combinaison est, pour l'essentiel, une simple somme réalisée par un additionneur 46. Ce pourrait également être une somme pondérée.

[0030] Le coefficient de surestimation $\alpha'_{n,i}$ est égal au rapport entre la somme $\hat{B}_{n,i} + \Delta B_{n,i}^{\max}$ délivrée par l'additionneur 46 et l'estimation à long terme retardée $\hat{B}_{n-\tau 3,i}$ (diviseur 47), plafonné à une valeur limitée $\alpha_{\max}$, par exemple $\alpha_{\max}=4$ (bloc 48). Le retard $\tau 3$ sert à corriger le cas échéant, dans les phases de montée ($\delta=1$), la valeur du coefficient de surestimation $\alpha'_{n,i}$ avant que les estimations à long terme aient été corrigées par les étapes 40 et 41 de la figure 3 (par exemple $\tau 3=3$).

[0031] L'estimation majorée $B'_{n,i}$ est finalement prise égale à $\alpha'_{n,i} \cdot \hat{B}_{n-\tau 3,i}$ (multiplieur 49).

[0032] La mesure $\Delta B_{n,i}^{\max}$ de la variabilité du bruit reflète la variance de l'estimateur de bruit. Elle est obtenue en fonction des valeurs de $S_{n,i}^i$ et de $B_{n,i}$ calculées pour un certain nombre de trames précédentes sur lesquelles le signal de parole ne présente pas d'activité vocale dans la bande i . C'est une fonction des écarts $|S_{n-k,i} - B_{n-k,i}|$ calculés pour un nombre K de trames de silence ($n-k \leq n$). Dans l'exemple représenté, cette fonction est simplement le maximum (bloc 50). Pour chaque trame n , le degré d'activité vocale $\gamma_{n,i}$ est comparé à un seuil (bloc 51) pour décider si l'écart $|S_{n,i} - B_{n,i}|$, calculé en 52-53, doit ou non être chargé dans une file d'attente 54 de K emplacements organisée en mode premier entré-premier sorti (FIFO). Si $\gamma_{n,i}$ ne dépasse pas le seuil (qui peut être égal à 0 si la fonction $g()$ a la forme de la figure 5), la FIFO 54 n'est pas alimentée, tandis qu'elle l'est dans le cas contraire. La valeur maximale contenue dans la FIFO 54 est alors fournie comme mesure de variabilité $\Delta B_{n,i}^{\max}$.

[0033] La mesure de variabilité $\Delta B_{n,i}^{\max}$ peut, en variante, être obtenue en fonction des valeurs $S_{n,f}$ (et non $S_{n,i}$) et $\hat{B}_{n,i}$. On procède alors de la même manière, sauf que la FIFO 54 contient non pas $|S_{n-k,i} - B_{n-k,i}|$ pour chacune des bandes i , mais plutôt

$$\max_{f \in \{f(1-1), f(1)\}} |S_{n-k,f} - \hat{B}_{n-k,i}|$$

[0034] Grâce aux estimations indépendantes des fluctuations à long terme du bruit $\hat{B}_{n,i}$ et de sa variabilité à court terme $\Delta B_{n,i}^{\max}$, l'estimateur majoré $B'_{n,i}$ procure une excellente robustesse aux bruits musicaux du procédé de débruitage.

[0035] Une première phase de la soustraction spectrale est réalisée par le module 55 représenté sur la figure 1. Cette phase fournit, avec la résolution des bandes i ($1 \leq i \leq I$), la réponse en fréquence H_i^1 d'un premier filtre de débruitage, en fonction des composantes $S_{n,i}$ et $B_{n,i}$ et des coefficients de surestimation $\alpha'_{n,i}$. Ce calcul peut être effectué pour chaque bande i selon la formule :

$$H_{n,i}^1 = \frac{\max\{S_{n,i} - \alpha'_{n,i} \cdot \hat{B}_{n,i}, \beta_i^1 \cdot \hat{B}_{n,i}\}}{S_{n-\tau 4,i}} \quad (7)$$

où $\tau 4$ est un retard entier déterminé tel que $\tau 4 \geq 0$ (par exemple $\tau 4=0$). Dans l'expression (7), le coefficient β_i^1 représente, comme le coefficient βp_i de la formule (3), un plancher servant classiquement à éviter les valeurs négatives ou trop faibles du signal débruité.

[0036] De façon connue (EP-A-0 534 837), le coefficient de surestimation $\alpha'_{n,i}$ pourrait être remplacé dans la formule (7) par un autre coefficient égal à une fonction de $\alpha'_{n,i}$ et d'une estimation du rapport signal-sur-bruit (par exemple $S_{n,i}/B_{n,i}$), cette fonction étant décroissante selon la valeur estimée du rapport signal-sur-bruit. Cette fonction est alors égale à $\alpha'_{n,i}$ pour les valeurs les plus faibles du rapport signal-sur-bruit. En effet, lorsque le signal est très bruité, il n'est a priori pas utile de diminuer le facteur de surestimation. Avantagusement, cette fonction décroît vers zéro pour les valeurs les plus élevées du rapport signal/bruit. Ceci permet de protéger les zones les plus énergétiques du spectre, où le signal de parole est le plus significatif, la quantité soustraite du signal tendant alors vers zéro.

[0037] Cette stratégie peut être affinée en l'appliquant de manière sélective aux harmoniques de la fréquence tonale

(« pitch ») du signal de parole lorsque celui-ci présente une activité vocale.

[0038] Ainsi, dans la réalisation représentée sur la figure 1, une seconde phase de débruitage est réalisée par un module 56 de protection des harmoniques. Ce module calcule, avec la résolution de la transformée de Fourier, la réponse en fréquence $H_{n,f}^2$ d'un second filtre de débruitage en fonction des paramètres $H_{n,i}^1$, $\alpha'_{n,i}$, $B_{n,i}$, δ_n , $S_{n,i}$ et de la fréquence tonale $f_p = F_e / T_p$ calculée en dehors des phases de silence par un module d'analyse harmonique 57. En phase de silence ($\delta_n=0$), le module 56 n'est pas en service, c'est-à-dire que $H_{n,f}^2 = H_{n,i}^1$ pour chaque fréquence f d'une bande i . Le module 57 peut appliquer toute méthode connue d'analyse du signal de parole de la trame pour déterminer la période T_p , exprimée comme un nombre entier ou fractionnaire d'échantillons, par exemple une méthode de prédiction linéaire.

[0039] La protection apportée par le module 56 peut consister à effectuer, pour chaque fréquence f appartenant à une bande i :

$$\begin{cases} H_{n,f}^2 = 1 & \text{si} \quad \begin{cases} S_{n,i} - \alpha'_{n,i} \cdot \hat{B}_{n,i} > \beta_{n,i}^2 \cdot \hat{B}_{n,i} \\ \text{et } \exists \eta \text{ entier} / |f - \eta \cdot f_p| \leq \Delta f / 2 \end{cases} \\ H_{n,f}^2 = H_{n,i}^1 & \text{sinon} \end{cases} \quad (8)$$

[0040] $\Delta f = F_e / N$ représente la résolution spectrale de la transformée de Fourier. Lorsque $H_{n,i}^2 = 1$, la quantité soustraite de la composante $S_{n,i}$ sera nulle. Dans ce calcul, les coefficients de plancher $\beta_{n,i}^2$ (par exemple $\beta_{n,i}^2 = \beta_i^1$) expriment le fait que certaines harmoniques de la fréquence tonale f_p peuvent être masquées par du bruit, de sorte qu'il n'est pas utile de les protéger.

[0041] Cette stratégie de protection est de préférence appliquée pour chacune des fréquences les plus proches des harmoniques de f_p , c'est-à-dire pour η entier quelconque.

[0042] Si on désigne par δf_p la résolution fréquentielle avec laquelle le module d'analyse 57 produit la fréquence tonale estimée f_p , c'est-à-dire que la fréquence tonale réelle est comprise entre $f_p - \delta f_p / 2$ et $f_p + \delta f_p / 2$, alors l'écart entre la η -ième harmonique de la fréquence tonale réelle et son estimation $\eta \times f_p$ (condition (9)) peut aller jusqu'à $\pm \eta \times \delta f_p / 2$. Pour les valeurs élevées de η , cet écart peut être supérieur à la demi-résolution spectrale $\Delta f / 2$ de la transformée de Fourier. Pour tenir compte de cette incertitude et garantir la bonne protection des harmoniques de la fréquence tonale réelle, on peut protéger chacune des fréquences f de l'intervalle $[\eta \times f_p - \eta \times \delta f_p / 2, \eta \times f_p + \eta \times \delta f_p / 2]$, c'est-à-dire remplacer la condition (9) ci-dessus par :

$$\exists \eta \text{ entier} / |f - \eta \cdot f_p| \leq (\eta \cdot \delta f_p + \Delta f) / 2 \quad (9')$$

Cette façon de procéder (condition (9')) présente un intérêt particulier lorsque les valeurs de η peuvent être grandes, notamment dans le cas où le procédé est utilisé dans un système à bande élargie.

[0043] Pour chaque fréquence protégée, la réponse en fréquence corrigée $H_{n,f}^2$ peut être égale à 1 comme indiqué ci-dessus, ce qui correspond à la soustraction d'une quantité nulle dans le cadre de la soustraction spectrale, c'est-à-dire à une protection complète de la fréquence en question. Plus généralement, cette réponse en fréquence corrigée $H_{n,f}^2$ pourrait être prise égale à une valeur comprise entre 1 et $H_{n,i}^1$ selon le degré de protection souhaité, ce qui correspond à la soustraction d'une quantité inférieure à celle qui serait soustraite si la fréquence en question n'était pas protégée.

[0044] Les composantes spectrales $S_{n,f}^2$ d'un signal débruité sont calculées par un multiplieur 58 :

$$S_{n,f}^2 = H_{n,f}^2 \cdot S_{n,i} \quad (10)$$

[0045] Ce signal $S_{n,f}^2$ est fourni à un module 60 qui calcule, pour chaque trame n , une courbe de masquage en appliquant un modèle psychoacoustique de perception auditive par l'oreille humaine.

[0046] Le phénomène de masquage est un principe connu du fonctionnement de l'oreille humaine. Lorsque deux fréquences sont entendues simultanément, il est possible que l'une des deux ne soit plus audible. On dit alors qu'elle est masquée.

[0047] Il existe différentes méthodes pour calculer des courbes de masquage. On peut par exemple utiliser celle

développée par J.D. Johnston («Transform Coding of Audio Signals Using Perceptual Noise Criteria », IEEE Journal on Selected Area in Communications, Vol. 6, No. 2, février 1988). Dans cette méthode, on travaille dans l'échelle fréquentielle des barks. La courbe de masquage est vue comme la convolution de la fonction d'étalement spectral de la membrane basilaire dans le domaine bark avec le signal excitateur, constitué dans la présente application par le signal $S_{n,f}^2$. La fonction d'étalement spectral peut être modélisée de la manière représentée sur la figure 7. Pour chaque bande de bark, on calcule la contribution des bandes inférieures et supérieures convoluées par la fonction d'étalement de la membrane basilaire :

$$C_{n,q} = \sum_{q'=0}^{q-1} \frac{S_{n,q'}^2}{(10^{10/10})^{(q-q')}} + \sum_{q'=q+1}^Q \frac{S_{n,q'}^2}{(10^{25/10})^{(q'-q)}} \quad (11)$$

où les indices q et q' désignent les bandes de bark ($0 \leq q, q' \leq Q$), et $S_{n,q}^2$, représente la moyenne des composantes $S_{n,f}^2$ du signal excitateur débruité pour les fréquences discrètes f appartenant à la bande de bark q'.

[0048] Le seuil de masquage $M_{n,q}$ est obtenu par le module 60 pour chaque bande de bark q, selon la formule :

$$M_{n,q} = C_{n,q} / R_q \quad (12)$$

où R_q dépend du caractère plus ou moins voisé du signal. De façon connue, une forme possible de R_q est :

$$10 \cdot \log_{10}(R_q) = (A+q) \cdot \chi + B \cdot (1-\chi) \quad (13)$$

avec $A=14,5$ et $B=5,5$. χ désigne un degré de voisement du signal de parole, variant entre zéro (pas de voisement) et 1 (signal fortement voisé). Le paramètre χ peut être de la forme connue :

$$\chi = \min \left\{ \frac{SFM}{SFM_{\max}}, 1 \right\} \quad (12)$$

où SFM représente, en décibels, le rapport entre la moyenne arithmétique et la moyenne géométrique de l'énergie des bandes de bark, et $SFM_{\max} = -60$ dB.

[0049] Le système de débruitage comporte encore un module 62 qui corrige la réponse en fréquence du filtre de débruitage, en fonction de la courbe de masquage $M_{n,q}$ calculée par le module 60 et des estimations majorées $B'_{n,i}$ calculées par le module 45. Le module 62 décide du niveau de débruitage qui doit réellement être atteint.

[0050] En comparant l'enveloppe de l'estimation majorée du bruit avec l'enveloppe formée par les seuils de masquage $M_{n,q}$, on décide de ne débruiter le signal que dans la mesure où l'estimation majorée $B'_{n,i}$ dépasse la courbe de masquage. Ceci évite de supprimer inutilement du bruit masqué par de la parole.

[0051] La nouvelle réponse $H_{n,f}^3$, pour une fréquence f appartenant à la bande i définie par le module 12 et à la bande de bark q, dépend ainsi de l'écart relatif entre l'estimation majorée $B'_{n,i}$ de la composante spectrale correspondante du bruit et la courbe de masquage $M_{n,q}$, de la manière suivante :

$$H_{n,f}^3 = 1 - \left(1 - H_{n,f}^2 \right) \cdot \max \left\{ \frac{\hat{B}'_{n,i} - M_{n,q}}{\hat{B}'_{n,i}}, 0 \right\} \quad (14)$$

[0052] En d'autres termes, la quantité soustraite d'une composante spectrale $S_{n,f}$, dans le processus de soustraction spectrale ayant la réponse fréquentielle $H_{n,f}^3$, est sensiblement égale au minimum entre d'une part la quantité soustraite de cette composante spectrale dans le processus de soustraction spectrale ayant la réponse fréquentielle $H_{n,f}^2$, et d'autre part la fraction de l'estimation majorée $B'_{n,i}$ de la composante spectrale correspondante du bruit qui, le cas

échéant, dépasse la courbe de masquage $M_{n,q}$.

[0053] La figure 8 illustre le principe de la correction appliquée par le module 62. Elle montre schématiquement un exemple de courbe de masquage $M_{n,q}$ calculée sur la base des composantes spectrales $S_{n,f}^2$ du signal débruité, ainsi que l'estimation majorée $B'_{n,i}$ du spectre du bruit. La quantité finalement soustraite des composantes $S_{n,f}$ sera celle représentée par les zones hachurées, c'est-à-dire limitée à la fraction de l'estimation majorée $B'_{n,i}$ des composantes spectrales du bruit qui dépasse la courbe de masquage.

[0054] Cette soustraction est effectuée en multipliant la réponse fréquentielle $H_{n,f}^3$ du filtre de débruitage par les composantes spectrales $S_{n,f}$ du signal de parole (multiplieur 64). Un module 65 reconstruit alors le signal débruité dans le domaine temporel, en opérant la transformée de Fourier rapide inverse (TFRI) inverse des échantillons de fréquence $S_{n,f}^3$ délivrés par le multiplieur 64. Pour chaque trame, seuls les $N/2=128$ premiers échantillons du signal produit par le module 65 sont délivrés comme signal débruité final s^3 , après reconstruction par addition-recouvrement avec les $N/2=128$ derniers échantillons de la trame précédente (module 66).

Revendications

1. Procédé de détection d'activité vocale dans un signal de parole numérique (s) traité par trames successives, dans lequel on soumet le signal de parole à un débruitage en tenant compte d'estimations du bruit compris dans le signal, mises à jour pour chaque trame d'une manière dépendant d'au moins un degré d'activité vocale ($\gamma_{n,i}$) déterminé pour ladite trame, **caractérisé en ce qu'on** procède à un débruitage a priori du signal de parole de chaque trame sur la base d'estimations du bruit obtenues lors du traitement d'au moins une trame précédente, et on analyse les variations d'énergie du signal débruité a priori pour détecter le degré d'activité vocale de ladite trame.
2. Procédé selon la revendication 1, dans lequel le degré d'activité vocale ($\gamma_{n,i}$) est un paramètre non binaire.
3. Procédé selon la revendication 2, dans lequel le degré d'activité vocale ($\gamma_{n,i}$) est une fonction, variant continûment entre 0 et 1.
4. Procédé selon l'une quelconque des revendications précédentes, dans lequel les estimations du bruit sont obtenues dans différentes bandes fréquentielles du signal, le débruitage a priori est effectué bande par bande, et il est déterminé un degré d'activité vocale ($\gamma_{n,i}$) pour chaque bande.
5. Procédé selon l'une quelconque des revendications précédentes, dans lequel on obtient une estimation du bruit $B_{n,i}$ pour la trame n dans une bande de fréquences i sous la forme :

$$\hat{B}_{n,i} = \gamma_{n,i} \cdot \hat{B}_{n-1,i} + (1-\gamma_{n,i}) \cdot \tilde{B}_{n,i}$$

avec

$$\tilde{B}_{n,i} = \lambda_B \cdot \hat{B}_{n-1,i} + (1-\lambda_B) \cdot S_{n,i}$$

où λ_B est un facteur d'oubli compris entre 0 et 1, $\gamma_{n,i}$ est le degré d'activité vocale déterminé pour la trame n dans la bande de fréquences i, et $S_{n,i}$ est une moyenne de l'amplitude du spectre du signal de parole de la trame n sur la bande i.

6. Procédé selon la revendication 5, dans lequel le signal débruité a priori $\hat{E}p_{n,i}$ relativement à une trame n et à une bande de fréquences i est de la forme :

$$\hat{E}p_{n,i} = \max \left\{ H_{p,i} \cdot S_{n,i}, \beta_{p,i} \cdot \hat{B}_{n-1,i} \right\}$$

où

$$Hp_{n,i} = \frac{S_{n,i} - \alpha'_{n-\tau_1,i} \cdot \hat{B}_{n-\tau_1,i}}{S_{n-\tau_2,i}}$$

, τ_1 est un entier au moins égal à 1, τ_2 est un entier au moins égal à 0, $\alpha'_{n-\tau_1,i}$ est un coefficient de surestimation déterminé pour la trame $n-\tau_1$ et la bande i , et β_{p_i} est un coefficient positif.

7. Procédé selon l'une quelconque des revendications précédentes, dans lequel on calcule une estimation à long terme ($\hat{E}_{n,i}$) de l'énergie du signal débruité a priori ($\hat{E}p_{n,i}$), et on compare cette estimation à long terme à une estimation instantanée (ba) de cette énergie, calculée sur la trame en cours, pour obtenir le degré d'activité vocale ($\gamma_{n,i}$) de ladite trame.
8. Détecteur d'activité vocale, comprenant des moyens de traitement adaptés pour mettre en oeuvre un procédé selon l'une quelconque des revendications précédentes.

Patentansprüche

1. Verfahren zum Erfassen von Stimmaktivität in einem in aufeinanderfolgenden Blöcken behandelten digitalen Sprachsignal (s), bei dem das Sprachsignal einer Rauschunterdrückung unter Berücksichtigung von Schätzungen des im Signal enthaltenen Rauschens unterzogen wird, die für jeden Block auf eine Weise aktualisiert werden, die von zumindest einem für den betreffenden Block bestimmten Grad der Stimmaktivität ($\gamma_{n,i}$) abhängt, **dadurch gekennzeichnet, daß** eine apriorische Rauschunterdrückung des Sprachsignals eines jeden Blocks auf der Grundlage von Schätzungen des Rauschens durchgeführt wird, die bei der Behandlung von mindestens einem vorhergehenden Block erhalten wurden, und die Variationen der Energie des apriorisch rauschunterdrückten Signals analysiert werden, um den Grad der Stimmaktivität des Blocks zu erfassen.
2. Verfahren nach Anspruch 1, bei dem der Grad der Stimmaktivität ($\gamma_{n,i}$) ein nicht-binärer Parameter ist.
3. Verfahren nach Anspruch 2, bei dem der Grad der Stimmaktivität ($\gamma_{n,i}$) eine ständig zwischen 0 und 1 variierende Funktion ist.
4. Verfahren nach einem der vorhergehenden Ansprüche, bei dem die Schätzungen des Rauschens in verschiedenen Frequenzbändern des Signals erhalten werden, die apriorische Rauschunterdrückung Band für Band durchgeführt wird, und ein Grad der Stimmaktivität ($\gamma_{n,i}$) für jedes Band bestimmt wird.
5. Verfahren nach einem der vorhergehenden Ansprüche, bei dem eine Schätzung des Rauschens $\hat{B}_{n,i}$ für den Block n in einem Frequenzband i in der Form

$$\hat{B}_{n,i} = \gamma_{n,i} \cdot \hat{B}_{n-1,i} + (1 - \gamma_{n,i}) \cdot \tilde{B}_{n,i}$$

mit

$$\tilde{B}_{n,i} = \lambda_B \cdot \hat{B}_{n-1,i} + (1 - \lambda_B) \cdot S_{n,i}$$

erhalten wird,

wobei λ_B ein zwischen 0 und 1 liegender Vergessensfaktor ist, $\gamma_{n,i}$ der für den Block n im Frequenzband i bestimmte Grad der Stimmaktivität ist, und $S_{n,i}$ ein Mittelwert der Amplitude des Spektrums des Sprachsignals des Blocks n im Band i ist.

6. Verfahren nach Anspruch 5, bei dem das apriorisch rauschunterdrückte Signal $\hat{E}p_{n,i}$ bezüglich eines Blocks n und eines Frequenzbandes i die Form aufweist:

$$\hat{E}p_{n,i} = \max\{Hp_{n,i} \cdot S_{n,i}, \beta p_i \cdot \hat{B}_{n-\tau_1,i}\}$$

wobei

$$Hp_{n,i} = \frac{S_{n,i} - \alpha'_{n-\tau_1,i} \cdot \hat{B}_{n-\tau_1,i}}{S_{n-\tau_2,i}}$$

, τ_1 eine ganze Zahl von mindestens gleich 1 ist, τ_2 eine ganze Zahl von mindestens gleich 0 ist, $\alpha'_{n-\tau_1,i}$ ein für den Block $n-\tau_1$ und das Band i bestimmter Überbewertungskoeffizient ist, und βp_i ein positiver Koeffizient ist.

7. Verfahren nach einem der vorhergehenden Ansprüche, bei dem eine Langzeitschätzung ($\bar{E}_{n,i}$) der Energie des apriorisch rauschunterdrückten Signals ($\hat{E}p_{n,i}$) berechnet wird und diese Langzeitschätzung mit einer an dem betreffenden Block berechneten, momentanen Schätzung (ba) dieser Energie verglichen wird, um den Grad der Stimmaktivität ($\gamma_{n,i}$) des Blocks zu erhalten.
8. Stimmaktivität-Erfassungseinrichtung mit einer zum Durchführen eines Verfahrens nach einem der vorhergehenden Ansprüche konzipierten Behandlungseinrichtung.

Claims

1. Method of detecting vocal activity in a digital speech signal (s) processed by successive frames, comprising the step of subjecting the speech signal to noise suppression taking account of estimates of the noise included in the signal, updated for each frame in a manner depending on at least one degree of vocal activity ($\gamma_{n,i}$) determined for said frame, **characterized in that** a priori noise suppression is applied to the speech signal of each frame on the basis of estimates of the noise obtained on processing at least one preceding frame, and energy variations of the a priori noise-suppressed signal are analyzed to detect the degree of vocal activity of said frame.
2. Method according to claim 1, wherein the degree of vocal activity ($\gamma_{n,i}$) is a non-binary parameter.
3. Method according to claim 2, wherein the degree of vocal activity ($\gamma_{n,i}$) is a function which varies in a continuous manner in the range from 0 to 1.
4. Method according to any one of the preceding claims, wherein the estimates of the noise are obtained in different frequency bands of the signal, the a priori noise suppression is effected band by band, and a degree of vocal activity ($\gamma_{n,i}$) is determined for each band.
5. Method according to any one of the preceding claims, wherein an estimate of the noise $\hat{B}_{n,i}$ is obtained for the frame n in a band of frequencies i in the form:

$$\hat{B}_{n,i} = \gamma_{n,i} \cdot \hat{B}_{n-1,i} + (1-\gamma_{n,i}) \cdot \tilde{B}_{n,i}$$

where

$$\tilde{B}_{n,i} = \lambda_B \cdot \hat{B}_{n-1,i} + (1-\lambda_B) \cdot S_{n,i}$$

where λ_B is a forgetting factor in the range from 0 to 1, $\gamma_{n,i}$ is the degree of vocal activity determined for the frame n in the band of frequencies i , and $S_{n,i}$ is an average speech signal amplitude in frame n in band i .

6. Method according to claim 5, in which the a priori noise-suppressed signal $\hat{E}p_{n,i}$ relative to a frame n and a band of frequencies i is of the form:

$$\hat{E}_{p_{n,i}} = \max \left\{ H_{p_{n,i}} \cdot S_{n,i} , \beta_{p_i} \cdot \hat{B}_{n-\tau_1,i} \right\}$$

where

$$H_{p_{n,i}} = \frac{S_{n,i} - \alpha'_{n-\tau_1,i} \cdot \hat{B}_{n-\tau_1,i}}{S_{n-\tau_2,i}}$$

, τ_1 is an integer at least equal to 1, τ_2 is an integer at least equal to 0, $\alpha'_{n-\tau_1,i}$ is an overestimation coefficient determined for the frame $n-\tau_1$ and the band i , and β_{p_i} is a positive coefficient.

7. Method according to any one of the preceding claims, wherein a long-term estimate ($\bar{E}_{n,i}$) of the energy of the a priori noise-suppressed signal ($\hat{E}_{p_{n,i}}$) is computed and said long-term estimate is compared with an instantaneous estimate (ba) of said energy, computed over the current frame, to obtain the degree of vocal activity ($\gamma_{n,i}$) of said frame.
8. A vocal activity detector, comprising processing means adapted to implement a method according to any one of the preceding claims.

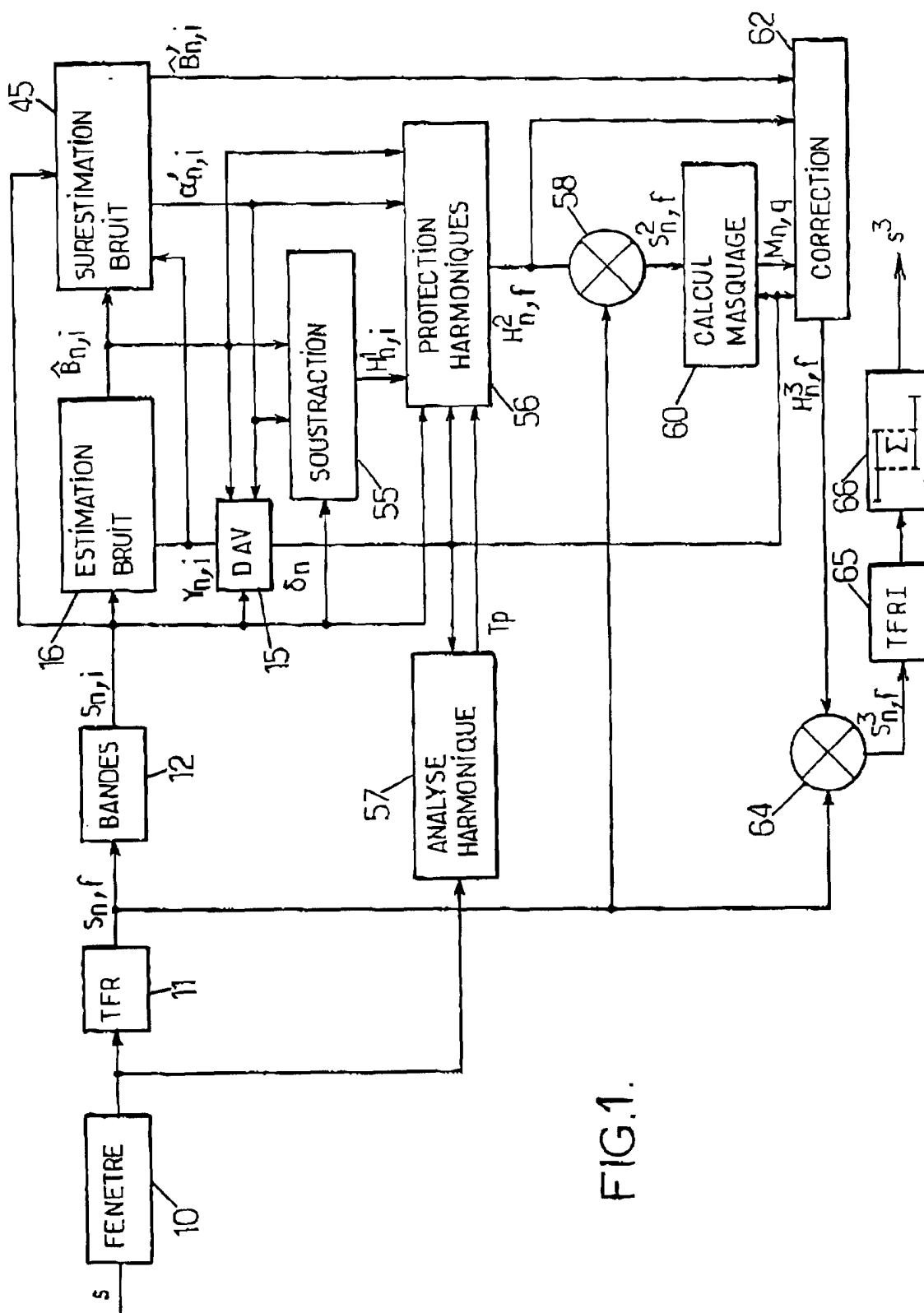


FIG.1.

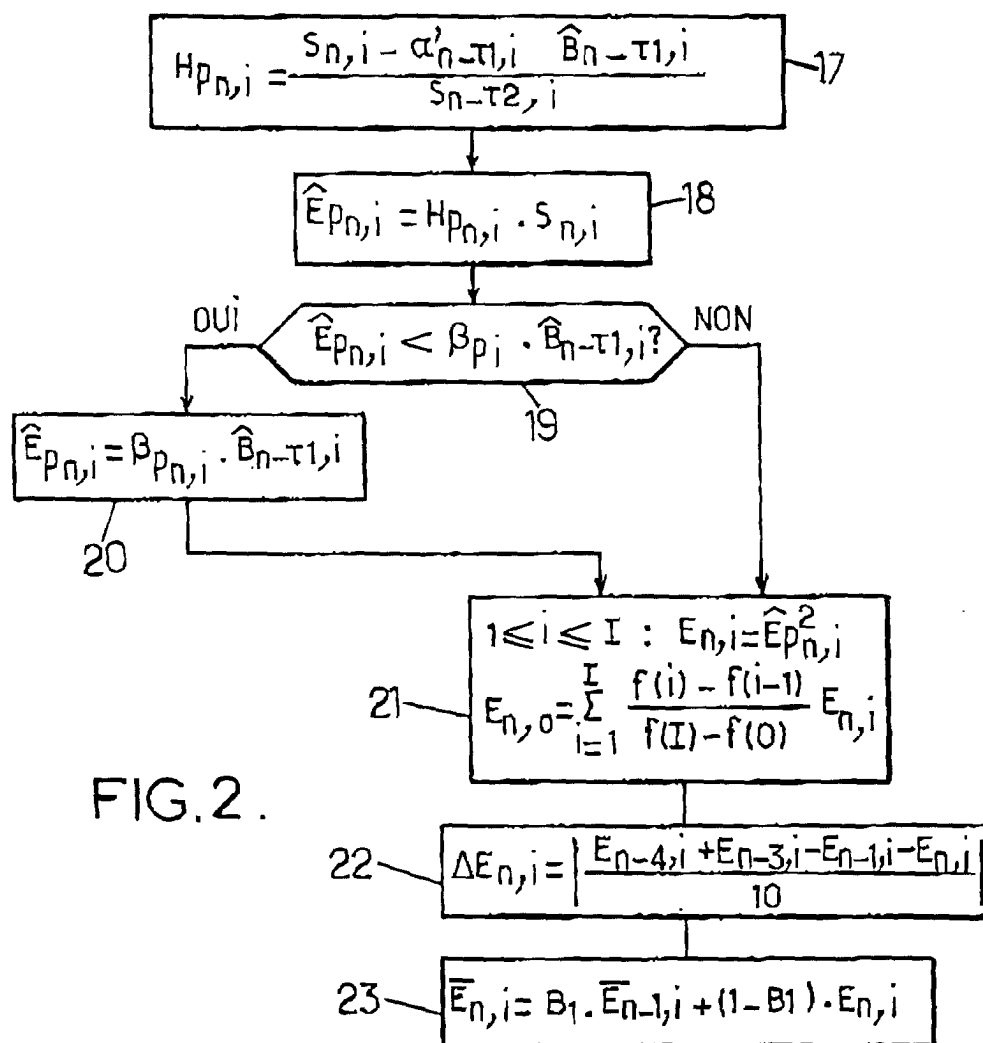


FIG. 4.

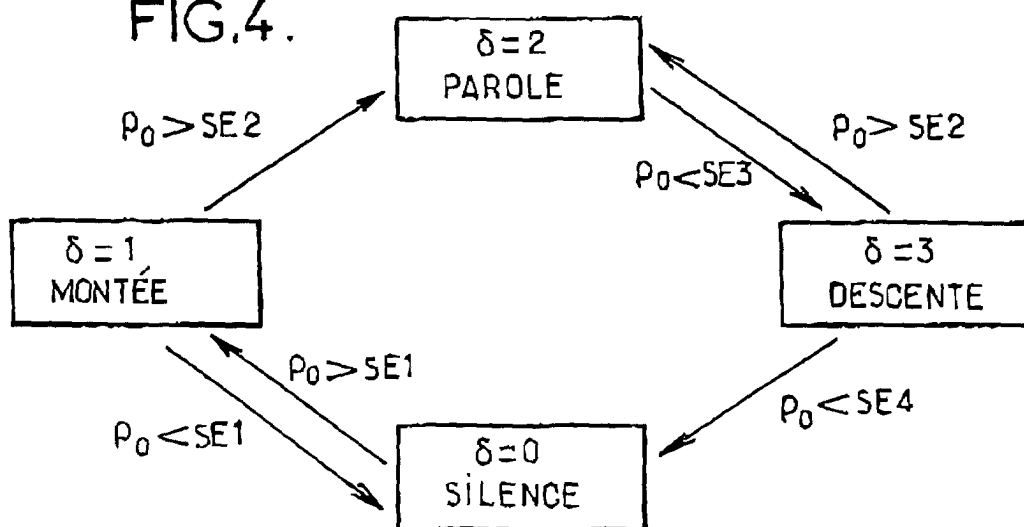
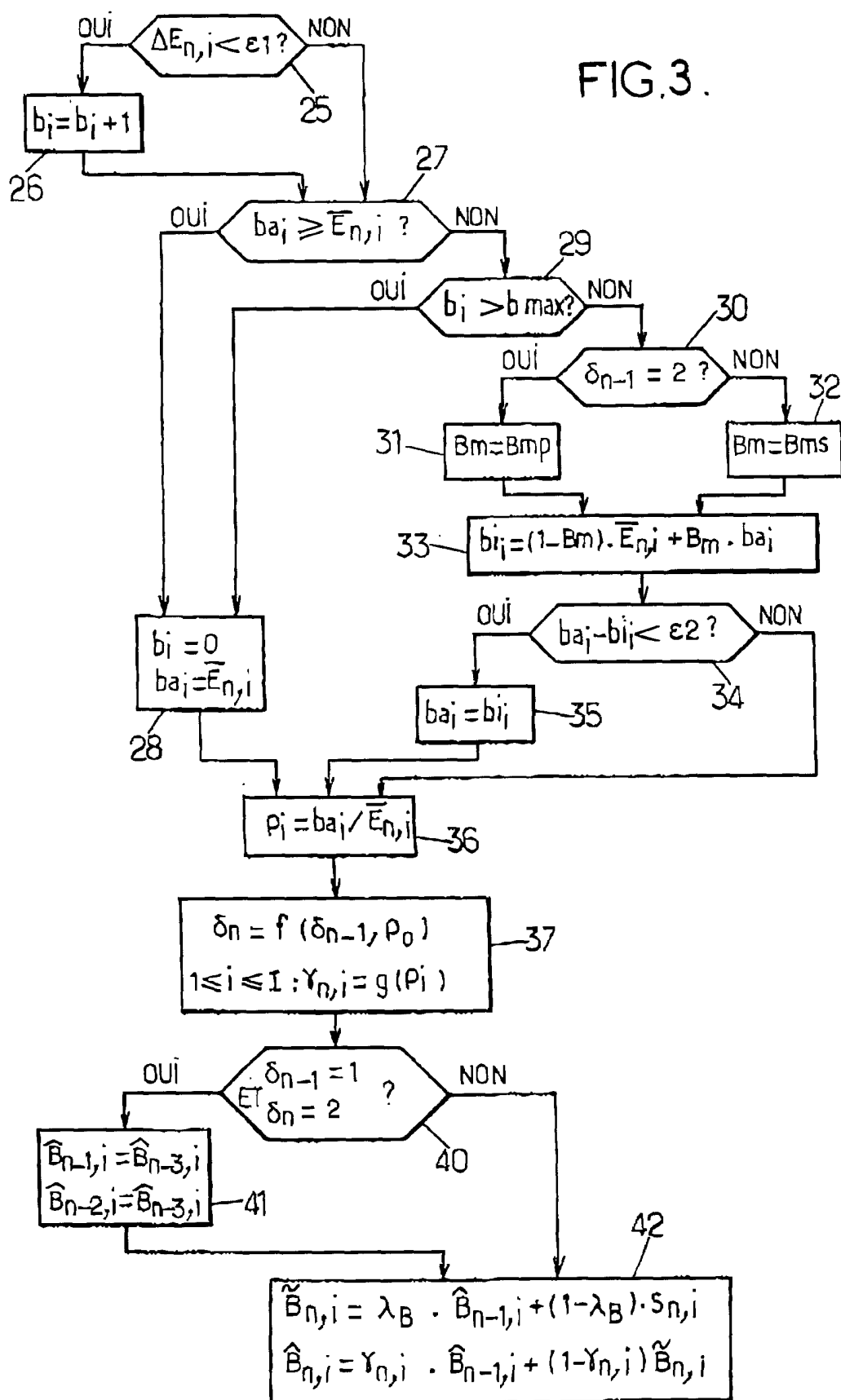


FIG.3.



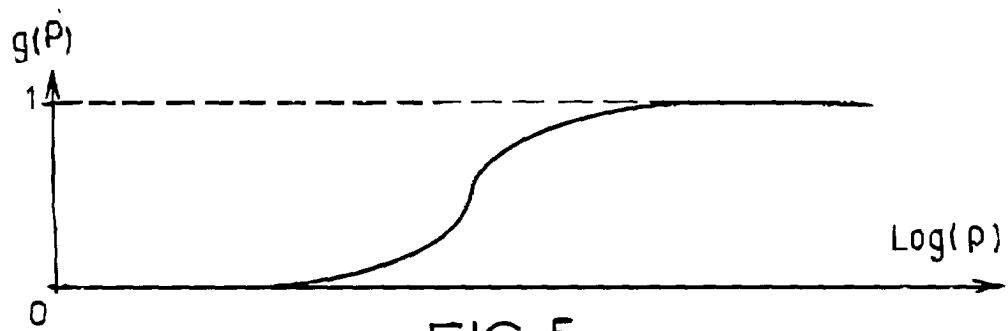


FIG. 5.

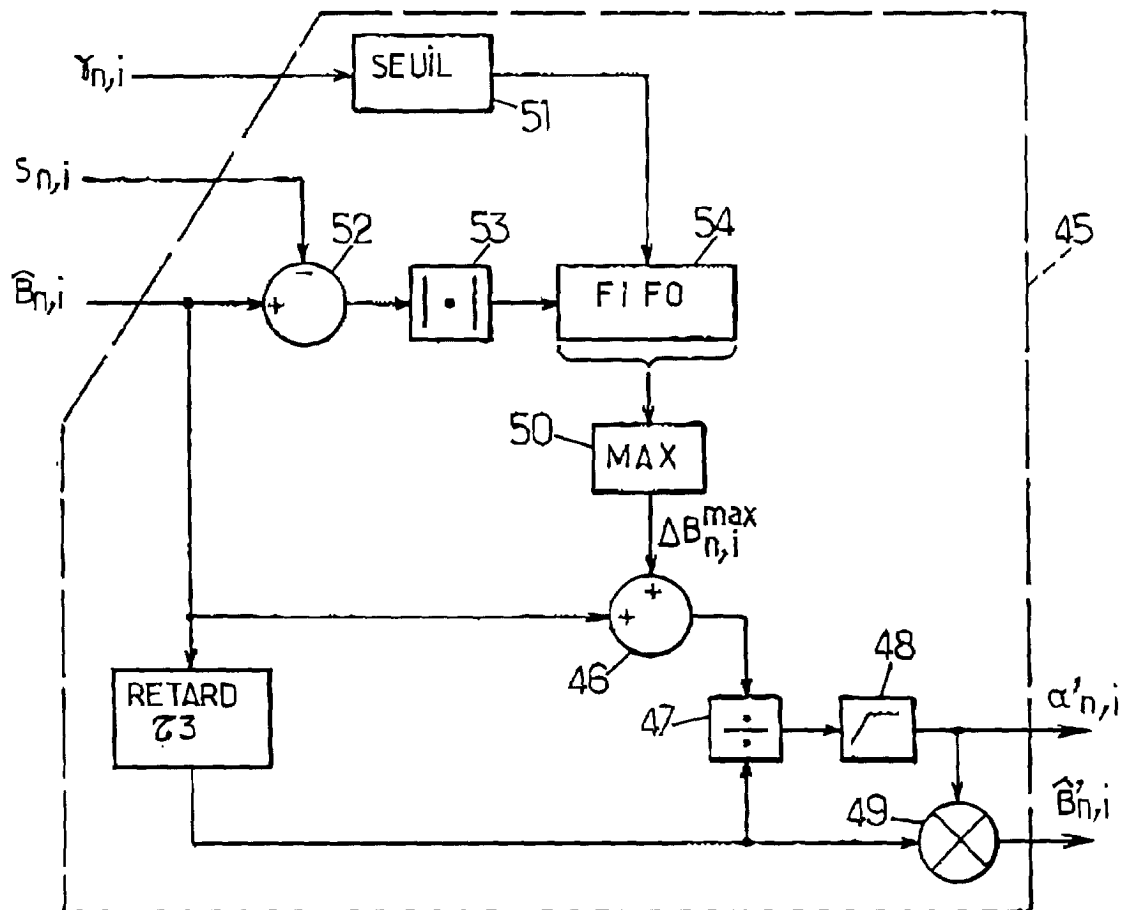


FIG. 6.

FIG. 7.

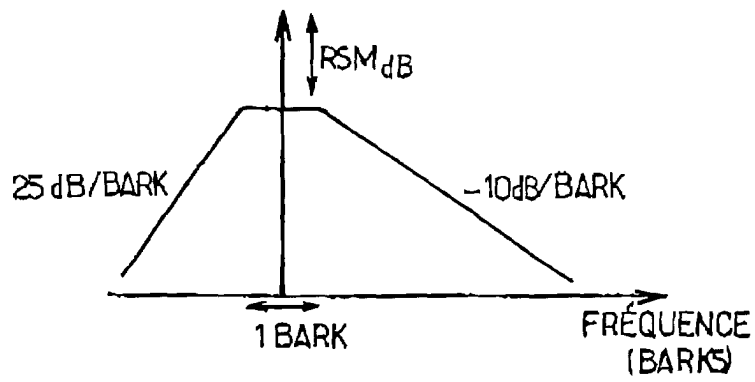


FIG. 8.

