

(19)



(11)

EP 1 944 755 A1

(12)

DEMANDE DE BREVET EUROPEEN

(43) Date de publication:

16.07.2008 Bulletin 2008/29

(51) Int Cl.:

G10L 13/06 ^(2006.01)**G10L 19/08** ^(2006.01)**G10L 21/04** ^(2006.01)**G10L 21/00** ^(2006.01)(21) Numéro de dépôt: **08100453.3**(22) Date de dépôt: **14.01.2008**

(84) Etats contractants désignés:

**AT BE BG CH CY CZ DE DK EE ES FI FR GB GR
HR HU IE IS IT LI LT LU LV MC MT NL NO PL PT
RO SE SI SK TR**

Etats d'extension désignés:

AL BA MK RS(30) Priorité: **15.01.2007 FR 0700257**(71) Demandeur: **France Télécom****75015 Paris (FR)**

(72) Inventeurs:

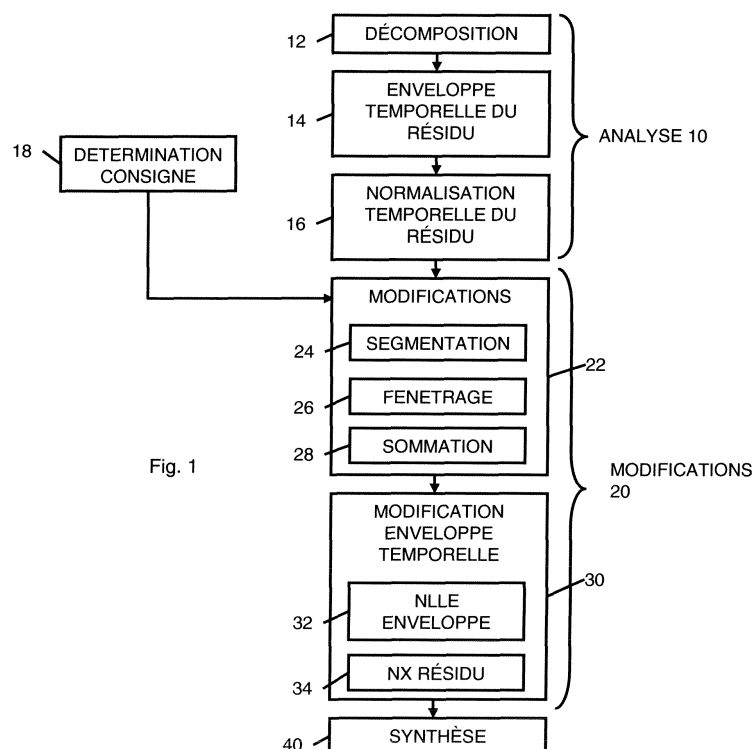
• **Rosec, Olivier**
22300 LANNION (FR)• **Vincent, Damien**
22300 LANNION (FR)(74) Mandataire: **Cabinet Plasseraud****52 rue de la Victoire****75440 Paris Cedex 09 (FR)**(54) **Modification d'un signal de parole**

(57) Ce procédé de modification des caractéristiques acoustiques d'un signal de parole comporte :

- une décomposition (12) du signal en une partie paramétrique et un résidu non paramétrique ;
- une estimation (14) de l'enveloppe temporelle du résidu ;
- une modification (22) de caractéristiques acoustiques

de la partie paramétrique et du résidu selon des consignes de modification ;

- une détermination (30), selon lesdites consignes de modification, d'une nouvelle enveloppe temporelle pour le résidu modifié ; et
- une synthèse (40) d'un signal de parole modifié à partir de la partie paramétrique modifiée et du résidu tel que modifié et avec la nouvelle enveloppe temporelle.



Description

[0001] La présente invention concerne la modification de la parole et plus particulièrement, la modification des paramètres acoustiques de signaux de parole décomposés en une partie paramétrique et une partie non paramétrique.

[0002] Il est connu de décomposer les signaux de parole selon des modèles dits « filtre - excitation ». Dans ces modèles, la parole est considérée comme une excitation glottique transformée par un filtre représentant le canal vocal.

[0003] L'excitation est obtenue par un filtrage inverse du signal de parole. Elle comprend parfois une partie qui est également paramétrique et un résidu. Le résidu correspond à la différence entre l'excitation et la modélisation paramétrique correspondante.

[0004] Lors de la modification des signaux de parole, les informations de fréquence, de rythme ou de timbre sont modifiées par le biais des paramètres du modèle.

[0005] Toutefois, ces modifications entraînent des distorsions audibles, notamment du fait d'un manque de contrôle de la cohérence temporelle, en particulier lors des modifications de fréquence fondamentale ou de timbre.

[0006] Par exemple, le document "Applying the Harmonic plus Noise Model in concatenative speech synthesis", IEEE Transactions on Speech and Audio Processing, vol. 9 (1), pp. 21-29, January 2001 de Y. Stylianou, prévoit d'utiliser un modèle harmonique plus bruit, ou modèle NNM, avec une modulation temporelle de la partie bruitée de façon à ce qu'elle s'intègre de manière naturelle à la partie déterministe. Cependant, cette méthode ne préserve pas la cohérence temporelle de la partie déterministe.

[0007] Une autre approche consiste à disposer d'un modèle de la source glottique suffisamment compact pour que l'allure du signal glottique puisse être maîtrisée lors de modifications du signal. Une telle approche est décrite par exemple dans le document "Toward a high-quality singing synthesizer with vocal texture control", Stanford University, 2002 de H. L. Lu. Néanmoins, un tel modèle ne capture pas toute l'information du signal glottique. Une information résiduelle doit être conservée et sa modification soulève le problème de manque de cohérence temporelle évoqué plus haut.

[0008] Dans le document "Time-scale modification of complex acoustic signals", ICASSP 1993, vol. 1, pp. 213-216, 1993 de T. F. Quatieri, R. B. Dunn and T. E. Hanna, il est proposé une méthode de modification de signaux de parole visant à préserver à la fois l'enveloppe spectrale et l'enveloppe temporelle. Cette méthode est appliquée uniquement à la modification de la durée de signaux acoustiques et n'est pas pratique dans la mesure où il n'est théoriquement pas possible de garantir l'existence d'une solution satisfaisant simultanément à ces deux propriétés. De surcroît, il n'existe pas de résultat de convergence de l'algorithme proposé et par conséquent, cette méthode ne permet pas d'avoir un contrôle suffisant sur les caractéristiques du signal résultant.

[0009] Ainsi, il n'existe pas de technique permettant de modifier les signaux de parole tout en assurant une bonne cohérence au niveau temporel.

[0010] Un des objectifs de la présente invention est de permettre une telle modification.

[0011] A cet effet, la présente invention a pour objet un procédé de modification des caractéristiques acoustiques d'un signal de parole, caractérisé en ce qu'il comporte :

- une décomposition du signal en une partie paramétrique et un résidu non paramétrique ;
- une estimation de l'enveloppe temporelle du résidu ;
- une modification de caractéristiques acoustiques de la partie paramétrique et du résidu selon des consignes de modification ;
- une détermination, selon lesdites consignes de modification, d'une nouvelle enveloppe temporelle pour le résidu modifié ; et
- une synthèse d'un signal de parole modifié à partir de la partie paramétrique modifiée et du résidu tel que modifié et avec la nouvelle enveloppe temporelle.

[0012] Grâce au traitement spécifique effectué sur les caractéristiques temporelles du résidu, la cohérence temporelle du signal modifié est améliorée.

[0013] Dans un mode de réalisation de l'invention, ladite décomposition du signal est une décomposition selon un modèle de type excitation - filtre. Une telle décomposition permet d'obtenir un résidu correspondant à une excitation glottique.

[0014] Avantagusement, l'estimation de l'enveloppe temporelle du résidu comprend l'estimation d'une première enveloppe, puis un lissage temporel de cette première enveloppe. Ce mode de réalisation permet d'obtenir une meilleure estimation de l'enveloppe temporelle.

[0015] Dans un mode de réalisation particulier, le procédé comprend en outre une normalisation temporelle du résidu en fonction de l'estimation de l'enveloppe temporelle. Ceci permet d'obtenir une expression du résidu sensiblement indépendante de ses caractéristiques temporelles.

[0016] Dans un mode de réalisation particulier, la normalisation temporelle du résidu comprend la division du résidu

par l'estimation de l'enveloppe temporelle.

[0017] Dans un autre mode de réalisation, la détermination d'une nouvelle enveloppe temporelle pour le résidu comprend une modification de paramètres de l'enveloppe temporelle du résidu selon lesdites consignes de modification et une application de l'enveloppe temporelle modifiée au résidu normalisé.

[0018] Dans un mode de réalisation, l'estimation de l'enveloppe temporelle et la détermination d'une nouvelle enveloppe temporelle sont confondues.

[0019] Avantagusement, la modification de caractéristiques acoustiques comprend une modification d'informations de fréquence fondamentale et de durée de la partie paramétrique et du résidu.

[0020] En outre, l'invention porte également sur un programme de mise en oeuvre du procédé décrit précédemment et un dispositif correspondant.

[0021] L'invention sera mieux comprise à la lumière de la description faite à titre d'exemple et en référence aux figures sur lesquelles :

- la figure 1 représente de manière générale un organigramme du procédé de l'invention ; et
- les figures 2A à 2D représentent différents stades de traitement d'un signal de parole.

[0022] Le procédé représenté en référence à la figure 1 débute par une étape 10 d'analyse du signal de parole qui comprend une décomposition 12 selon un modèle excitation - filtre, c'est-à-dire une décomposition du signal de parole en une partie paramétrique et une partie non paramétrique, appelée résidu et correspondant à une partie de l'excitation glottique.

[0023] Une pratique courante pour la mise en oeuvre de l'étape 12 est l'utilisation de techniques de prédiction linéaire telles que celles décrites dans le document de J. Makhoul, "Linear Prediction: a tutorial review", Proceedings of the IEEE, vol. 63(4), pp. 561-580, Avril 1975.

[0024] Dans le mode de réalisation décrit en exemple, la décomposition 12 du signal de parole $s(n)$ est réalisée à l'aide d'une auto-régression, ou modèle AR, de la forme suivante :

$$s(n) = \sum_{k=1}^p a_k s(n-k) + e(n).$$

[0025] Dans cette équation, les termes a_k désignent les coefficients d'un filtre de type AR modélisant le conduit vocal et le terme $e(n)$ est le signal résiduel relatif à la partie excitation, avec n un indice de trame de signal. Notons que si l'ordre du modèle est suffisamment grand alors $e(n)$ n'est pas corrélé à $s(n)$.

[0026] Cela s'écrit formellement $E[e(n)s(n-m)] = 0$ pour tout entier m , où $E[.]$ désigne l'espérance mathématique.

[0027] En pratique, des ordres typiques de 10 et de 16 sont choisis pour des signaux de parole échantillonnés respectivement à 8 et 16 kHz.

[0028] En multipliant l'équation précédente à gauche et à droite par $s(n-m)$ et en passant à l'espérance mathématique, on aboutit aux équations de Yule-Walker définies par :

$$r(m) = - \sum_{k=1}^p a_k r(m-k)$$

où r est la fonction d'autocorrélation définie par : $r(m) = E[s(n)s(n-m)]$.

[0029] Un estimateur de $r(m)$ est donné par :

$$r(m) = \frac{1}{N-p} \sum_{n=1}^{N-p} s(n)s(n-m).$$

[0030] En pratique, seules les $p+1$ premières valeurs de la fonction d'autocorrélation sont nécessaires pour l'estimation des coefficients de filtrage a_k . L'expression de cette dernière équation sous forme matricielle conduit à la résolution du système linéaire suivant :

5

$$\begin{bmatrix} r(0) & r(1) & \cdots & r(p-1) \\ r(1) & r(0) & \cdots & r(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ r(p-1) & r(p-2) & \cdots & r(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} r(1) \\ r(2) \\ \vdots \\ r(p) \end{bmatrix}.$$

10 **[0031]** Ainsi, l'estimation des coefficients revient à l'inversion d'une matrice de Toeplitz, ce qui peut être réalisé à l'aide de procédures classiques et notamment à l'aide de l'algorithme décrit par J. Durbin, "The fitting of time-series models", Rev. Inst. Int. Statistics.

[0032] En variante, la décomposition 12 permet d'obtenir, pour l'excitation, un modèle paramétrique en sus du résidu.

15 **[0033]** Par exemple, la décomposition excitation - filtre est réalisée en utilisant une information a priori sur l'excitation. Ainsi, l'excitation peut être modélisée en intégrant des informations liées au processus de production de la parole, notamment via un modèle paramétrique de la dérivée de l'onde de débit glottique (DODG), tel que par exemple le modèle LF proposé par Liljencrants et Fant dans "A four-parameter model of glottal flow", STL-QPSR, vol. 4, pp. 1-13, 1985. Ce modèle est entièrement défini par la donnée de la période fondamentale T_0 , de trois paramètres de forme qui sont un quotient ouvert de périodes, un coefficient d'asymétrie et un coefficient de phase de retour, d'un paramètre de position correspondant à l'instant de fermeture de glotte et d'un terme b_0 caractérisant l'amplitude de la DODG.

20 **[0034]** Dans ce contexte, le signal de parole peut être représenté par le modèle d'auto-régression exogène ARX-LF suivant :

25

$$s(n) = \sum_{k=1}^p a_k s(n-k) + b_0 u(n) + e(n)$$

où $u(n)$ désigne le signal correspondant au modèle LF de la DODG.

30 **[0035]** L'estimation simultanée des paramètres de la DODG et des paramètres liés au filtre est délicate, notamment parce que l'optimisation selon les paramètres de forme et de position est un problème non linéaire. Toutefois, lorsque T_0 et u sont fixés, l'optimisation selon les paramètres a_k et b_0 est un problème linéaire classique, pour lequel un estimateur des moindres carrés peut être obtenu analytiquement. Sur la base de constat, une méthode efficace a été proposée par D. Vincent, O. Rosec, et T. Chonavel, dans la publication "Estimation of LF glottal source parameters based on ARX model", Interspeech'05, pp. 333-336, Lisbonne, Portugal, 2005.

35 **[0036]** Dans ce mode de réalisation, à l'issue de la procédure d'estimation, le procédé délivre :

- des paramètres caractérisant complètement la DODG selon le modèle LF;
- des paramètres de filtre a_k ;
- 40 - le résidu $e(n)$ correspondant à l'erreur de modélisation liée au modèle ARX-LF.

[0037] De manière générale, à l'issue de l'étape 12, le procédé délivre une modélisation du signal de parole $s(n)$ sous la forme d'une partie paramétrique et d'un résidu qui est non paramétrique.

[0038] L'étape d'analyse 10 comprend ensuite une estimation 14 de l'enveloppe temporelle du résidu.

45 **[0039]** Dans le mode de réalisation décrit, l'enveloppe temporelle est définie comme le module du signal analytique et est obtenue par une transformation dite de Hilbert. Ainsi, l'enveloppe temporelle $d(t)$ du résidu $e(t)$ s'écrit :

50

$$d(t) = |x_e(t)| \text{ avec } x_e(t) = e(t) + iH(e(t)),$$

où H désigne l'opération de transformation de Hilbert.

[0040] Avantageusement, l'estimation 14 comprend un lissage de l'enveloppe temporelle du résidu. Cela procure une meilleure estimation notamment pour des sons voisés pour lesquels l'enveloppe est périodique de période T_0 , avec T_0 désignant l'inverse de la fréquence fondamentale f_0 . Par exemple, une modélisation cepstrale d'ordre K de l'enveloppe peut être utilisée. Celle-ci s'écrit sous la forme :

55

$$\ln(d(n)) = \frac{1}{2} \sum_{k=-K}^K c_k \exp(2i\pi k n f_0 / f_s) + \varepsilon(n)$$

[0041] L'estimation des coefficients cepstraux c_k se fait alors par minimisation de $\varepsilon(n)$ au sens des moindres carrés. Plus précisément, l'équation précédente s'écrit sous la forme matricielle suivante :

$$\mathbf{d} = \mathbf{M}\mathbf{c} + \varepsilon,$$

avec

$$\mathbf{d} = \frac{1}{2} [\ln(d(-N)), \dots, \ln(d(N))]^T$$

$M_{n+(N+1), k+(K+1)} = \exp(2i\pi k n f_0 / f_s)$, $n \in \{-N, \dots, N\}$, $k \in \{-K, \dots, K\}$
et

$$\mathbf{c} = [c_{-K}, \dots, c_K]^T.$$

[0042] Dans ces équations, l'exposant T représente l'opérateur de transposition.

[0043] La solution optimale au sens des moindres carrés est alors

$$\mathbf{c} = (\mathbf{M}^H \mathbf{M})^{-1} \mathbf{M}^H \mathbf{d}$$

où H désigne l'opérateur de transposition hermitienne. L'enveloppe correspondante s'écrit de la façon suivante :

$$\hat{d}(n) = \exp\left(\frac{1}{2} \sum_{k=-K}^K \hat{c}_k \exp(2i\pi k n f_0 / f_s)\right)$$

[0044] Une fois l'enveloppe temporelle du résidu estimée, le procédé comprend une étape 16 de normalisation temporelle du résidu. Dans ce document, normalisation temporelle signifie obtention d'un résidu sensiblement invariant au niveau temporel, plus précisément, obtention d'un résidu dont l'enveloppe temporelle est constante.

[0045] Dans le mode de réalisation décrit, l'étape 16 est mise en oeuvre en divisant le résidu par l'expression de l'enveloppe temporelle selon l'équation suivante :

$$\tilde{e}(n) = \frac{e(n)}{\hat{d}(n)}.$$

[0046] En parallèle de l'analyse 10, le procédé comprend une étape 18 de détermination de consignes de modification du signal de parole. Ces consignes peuvent être de deux types.

[0047] Dans un premier cas, une cible a été définie pour chacun des paramètres à modifier. Cela est notamment le cas en synthèse de la parole où de nombreux algorithmes de prédiction de la durée, de la fréquence fondamentale ou encore de l'énergie des signaux existent. Par exemple, des valeurs de fréquence fondamentale et d'énergie peuvent être estimées pour le début et la fin de chaque syllabe ou encore de chaque phonème de l'énoncé. De même, la durée de chaque syllabe ou de chaque phonème peut être prédite. Etant donné ces cibles numériques et le signal de parole,

des coefficients de modification peuvent être obtenus en faisant le rapport entre la mesure effectuée sur le signal et la valeur de la cible prédite correspondante.

[0048] Dans un deuxième cas, de telles cibles ne sont pas disponibles, mais il est possible de définir un ensemble de coefficients de modification pour la modification des paramètres désirés. Par exemple, un coefficient de modification de fréquence fondamentale de 0,5 permet de diviser par 2 la hauteur de voix perçue. Notons que ces coefficients de modification peuvent être définis de manière globale sur l'ensemble de l'énoncé ou de façon plus locale, par exemple à l'échelle d'une syllabe ou d'un mot.

[0049] Le procédé comprend ensuite une étape 20 de modification du signal de parole $s(n)$ selon les consignes déterminées précédemment.

[0050] Les modifications opérées concernent la fréquence fondamentale, la durée et l'énergie des signaux de parole. En outre, lorsqu'une analyse utilisant une DODG est mise en oeuvre étant donné qu'une décomposition de type source-filtre est disponible, des modifications des paramètres de la qualité vocale peuvent être opérées en altérant le quotient ouvert, le coefficient d'asymétrie, ou encore le coefficient de phase de retour.

[0051] L'étape 20 de modification comprend tout d'abord une modification 22 de la partie paramétrique du modèle correspondant au signal de parole et du résidu normalisé.

[0052] Dans le mode de réalisation décrit, cette modification vise la fréquence fondamentale ainsi que la durée et est mise en oeuvre de manière classique avec une technique connue sous le nom de TD-PSOLA (en anglais Time Domain Pitch Synchronous Overlap and Add) telle que décrite dans la publication "Non-parametric techniques for pitch-scale and time-scale modification of speech", Speech Communication, vol. 16, pp. 175-205, 1995 par E. Moulines et J. Laroche.

[0053] Cette technique permet d'opérer conjointement la modification de la durée et de la fréquence fondamentale avec les coefficients respectifs $\alpha(t)$ et $\beta(t)$.

[0054] En référence aux figures 2A à 2D, les principales étapes du fonctionnement de la technique TD-PSOLA sont illustrées.

[0055] La figure 2A représente le signal de parole à modifier $s(n)$. Au cours d'une étape 24, ce signal est segmenté en trames de manière dite pitch-synchrone, c'est-à-dire que chaque segment a une durée correspondant à l'inverse de la fréquence fondamentale du signal.

[0056] En effet, les instants de fermeture de glotte, aussi appelés instants d'analyse, sont situés au voisinage des maxima d'énergie du signal de parole et le traitement TD-PSOLA permet une bonne préservation des caractéristiques du signal de parole au voisinage des extrémités des segments obtenus par analyse pitch-synchrone. Ainsi, lorsque ces instants sont repérés avec une précision satisfaisante, les performances de TD-PSOLA sont optimisées. Une telle segmentation pitch-synchrone est obtenue, par exemple, par des techniques à base de délais de groupe ou encore à partir de la méthode proposée par D. Vincent, O. Rosec, et T. Chonavel, dans la publication "Glottal closure instant estimation using an appropriateness measure of the source and continuity constraints", IEEE ICASSP'06, vol. 1, pp. 381-384, Toulouse, France, Mai 2006.

[0057] Avantagusement, cette étape de marquage pitch-synchrone est réalisée hors-ligne, c'est-à-dire non en temps réel, ce qui permet de réduire la charge de calcul pour une mise en oeuvre en temps réel.

[0058] En fonction des facteurs de modification souhaités pour la fréquence fondamentale et la durée, les instants séparant les segments sont modifiés selon les règles suivantes:

- pour un allongement de durée, certains segments sont dupliqués afin d'augmenter artificiellement le nombre d'impulsions glottiques ;
- pour une réduction de la durée, certains segments sont supprimés;
- pour une augmentation de la fréquence fondamentale, c'est-à-dire un rendu plus aigu, les instants d'analyse sont rapprochés, ce qui nécessite éventuellement la duplication de segments pour conserver la durée totale ; et
- pour une diminution de la fréquence fondamentale, c'est-à-dire un rendu plus grave, les instants d'analyse sont écartés, ce qui nécessite éventuellement la suppression de segments pour conserver la durée totale.

[0059] Une description détaillée de ces règles se trouve dans la publication "Non-parametric techniques for pitch-scale and time-scale modification of speech", Speech Communication, vol. 16, pp. 175-205, 1995 par E. Moulines et J. Laroche.

[0060] A l'issue de cette étape, le signal comprend un nombre entier de segments ou trames, chacun d'une durée correspondant à une période qui est l'inverse de la fréquence fondamentale modifiée, comme cela est représenté sur la figure 2B.

[0061] Le traitement de modification comprend ensuite un fenêtrage 26 du signal autour des instants d'analyse, c'est-à-dire des instants séparant les segments. Au cours de ce fenêtrage, on sélectionne, pour chaque instant d'analyse, une portion du signal fenêtrée autour de cet instant. Cette portion de signal est appelée "signal court-terme" et s'étend, dans l'exemple, sur une durée correspondant à deux fois la période fondamentale modifiée comme représenté en référence à la figure 2C.

[0062] Le traitement de modification comprend enfin une sommation 28 des signaux court-terme qui sont recentrés sur les instants de synthèse et ajoutés comme représenté en référence à la figure 2D.

[0063] En variante, l'étape 22 est réalisée avec une technique de type HNM (en anglais *Harmonic plus Noise Model*), ou de type vocodeur de phase. Les modifications de fréquence fondamentale et de durée peuvent également être réalisées par des techniques différentes.

[0064] Dans la suite, le résidu normalisé modifié, c'est-à-dire le résidu normalisé dont les informations de fréquence fondamentale et/ou de durée ont été modifiées, est noté $\tilde{d}^{modif}(n)$.

[0065] Le procédé comprend ensuite une étape 30 de modification de l'enveloppe temporelle du résidu. Plus précisément, cette étape permet de substituer aux caractéristiques temporelles d'origine du résidu, des caractéristiques temporelles en accord avec les modifications souhaitées.

[0066] L'étape 30 débute par une détermination 32 de nouvelles caractéristiques temporelles du résidu. Dans l'exemple, il s'agit de la modification de l'enveloppe temporelle du résidu, telle qu'obtenue à l'issue de l'étape 14.

[0067] Comme indiqué précédemment, en considérant une trame pitch-synchrone du signal, deux types de modifications peuvent être opérées conjointement ou non :

- une modification de la fréquence fondamentale ; et
- une modification des paramètres liés à la qualité vocale.

[0068] La modification de la fréquence fondamentale consiste en une modification de l'enveloppe temporelle pour la rendre cohérente avec le résidu normalisé dont la fréquence fondamentale a été préalablement modifiée.

[0069] Un mode de réalisation d'une telle modification consiste en une dilatation/contraction de l'enveloppe temporelle originale $\hat{d}(n)$ afin d'en préserver la forme générale.

[0070] Etant donné la valeur de fréquence fondamentale modifiée f_0^{modif} , l'enveloppe temporelle modifiée d^{modif} s'écrit alors de la manière suivante :

$$d^{modif}(n) = \exp\left(\frac{1}{2} \sum_{k=-K}^K \hat{c}_k \exp(2i\pi kn f_0^{modif} / f_s)\right).$$

[0071] Lorsque des modifications des paramètres liés à la qualité vocale sont opérées, la forme de l'enveloppe temporelle doit être modifiée. Par exemple, lorsque des modifications du coefficient ouvert sont opérées, il convient d'appliquer des facteurs de dilatation/contraction différents sur respectivement les parties ouvertes et fermées du cycle glottique.

[0072] Par exemple, on effectue une modification du quotient ouvert de sorte que la durée de la phase ouverte devienne T_e^{modif} avec $T_e^{modif} < T_0$, avec T_0 qui est la longueur d'un cycle glottique dont l'instant de fermeture coïncide avec l'origine des temps et une phase ouverte originale de durée T_e . Dans ce cas, pour conserver la même période fondamentale, il convient de dilater le signal selon les coefficients suivants :

$$\alpha_1 = \frac{T_0 - T_e^{modif}}{T_0 - T_e} \text{ pour la phase fermée ;}$$

$$\alpha_2 = \frac{T_e^{modif}}{T_e} \text{ pour la phase ouverte.}$$

[0073] Mathématiquement, cela revient à déterminer une enveloppe temporelle de la forme suivante :

$$d^{modif}(t) = \exp\left(\frac{1}{2} \sum_{k=-K}^K \hat{c}_k \exp(2i\pi kg(t/T_0^{modif}))\right),$$

où la fonction g est définie par :

$$g(t) = \begin{cases} \frac{T_0 - T_e^{modif}}{T_0 - T_e} t & \text{pour } t \in [0, T_0 - T_e] \\ T_0 - T_e^{modif} + \frac{T_e^{modif}}{T_e} (t - (T_0 - T_e)) & \text{pour } t \in [T_0 - T_e, T_0] \end{cases}.$$

[0074] Bien entendu, d'autres types de modification de paramètres de la qualité vocale sont possibles selon des principes similaires.

[0075] L'étape 30 comprend ensuite une détermination 34 du nouveau résidu. Dans l'exemple, ce nouveau résidu s'obtient par multiplication du résidu $\tilde{e}^{modif}(n)$ par l'enveloppe modifiée d^{modif} .

[0076] Le résidu d'origine a donc été normalisé, modifié, puis combiné avec la nouvelle enveloppe temporelle. Ceci permet d'assurer la cohérence de son enveloppe temporelle avec les modifications de fréquence fondamentale et/ou de qualité vocale.

[0077] Dans le mode de réalisation décrit, l'excitation est confondue avec le résidu, ce qui correspond au cas où le résidu est obtenu par simple filtrage linéaire inverse et où l'excitation ne comporte par de partie paramétrique.

[0078] Dans le cas où l'excitation est composée d'une source glottique modélisable par un modèle paramétrique et d'un résidu, il convient d'opérer le même type de modification sur la source glottique ainsi paramétrée en ajustant les paramètres de fréquence fondamentale et de qualité vocale.

[0079] Le procédé comporte enfin une étape 40 de synthèse du signal modifié. Cette synthèse consiste en un filtrage du signal obtenu à l'issue de l'étape 20 par le filtre du conduit vocal tel que défini lors de l'étape 12. L'étape 40 comprend également une addition - recouvrement des trames ainsi filtrées. Cette étape de synthèse est classique et ne sera pas décrite plus en détails ici.

[0080] Ainsi, le traitement spécifique de l'enveloppe temporelle du résidu permet d'obtenir une modification assurant une bonne cohérence temporelle.

[0081] Bien entendu, d'autres modes de réalisation peuvent être envisagés.

[0082] Tout d'abord, le résidu peut être décomposé en sous-bandes. Dans ce cas, les étapes 14, 16 et 20 sont réalisées sur tout ou partie des sous-bandes considérées séparément. Le résidu final obtenu est alors la somme des résidus modifiés issus des différentes sous-bandes.

[0083] En outre, le résidu peut faire l'objet d'une décomposition en une partie déterministe et une partie stochastique. Dans ce cas, les étapes 14, 16 et 20 sont réalisées pour chacune des parties considérées. Là encore, le résidu final obtenu est alors la somme des composantes déterministes et stochastiques modifiées.

[0084] En outre, ces deux variantes peuvent être combinées, de sorte qu'un traitement séparé sur chaque sous-bande et pour chacune des composantes déterministe et stochastique peut être effectué.

[0085] Dans un autre mode de réalisation, les différentes étapes de l'invention peuvent être réalisées dans un ordre différent. Par exemple, l'enveloppe temporelle est modifiée avant que les modifications ne soient faites sur le signal. Ainsi, les modifications sont apportées sur le résidu avec sa nouvelle enveloppe temporelle et non sur le résidu normalisé comme dans l'exemple décrit précédemment.

[0086] Selon un autre mode de réalisation, les étapes de normalisation du résidu et de détermination de nouvelles caractéristiques temporelles sont combinées. Dans un tel mode de réalisation, le résidu est directement modifié par un facteur temporel déterminé à partir de son enveloppe temporelle et des consignes de modification. Ce facteur temporel permet à la fois de supprimer la dépendance du résidu avec ses caractéristiques temporelles d'origine et d'appliquer de nouvelles caractéristiques temporelles.

[0087] Par ailleurs, l'invention peut être mise en oeuvre par un programme contenant des instructions spécifiques qui, lorsqu'elles sont exécutées par un ordinateur, entraînent la réalisation des étapes décrites précédemment.

[0088] L'invention peut également être mise en oeuvre par un dispositif comportant des moyens appropriés, tels que des microprocesseurs, microcalculateurs et mémoires associées, ou encore des composants électroniques programmés.

[0089] Un tel dispositif peut être adapté pour mettre en oeuvre n'importe quel mode de réalisation du procédé décrit précédemment.

Revendications

1. Procédé de modification des caractéristiques acoustiques d'un signal de parole $(s(n))$ caractérisé en ce qu'il

comporte :

- une décomposition (12) du signal en une partie paramétrique et un résidu non paramétrique ($e(n)$) ;
- une estimation (14) de l'enveloppe temporelle du résidu ;
- une modification (22) de caractéristiques acoustiques de la partie paramétrique et du résidu selon des consignes de modification ;
- une détermination (30), selon lesdites consignes de modification, d'une nouvelle enveloppe temporelle pour le résidu modifié ; et
- une synthèse (40) d'un signal de parole modifié à partir de la partie paramétrique modifiée et du résidu tel que modifié et avec la nouvelle enveloppe temporelle.

2. Procédé selon la revendication 1, **caractérisé en ce que** ladite décomposition du signal est une décomposition selon un modèle de type excitation - filtre.

3. Procédé selon l'une quelconque des revendications 1 et 2, **caractérisé en ce que** l'estimation de l'enveloppe temporelle du résidu comprend l'estimation d'une première enveloppe puis un lissage temporel de cette première enveloppe.

4. Procédé selon l'une quelconque des revendications 1 à 3, **caractérisé en ce qu'il** comprend en outre une normalisation temporelle (16) du résidu en fonction de l'estimation de l'enveloppe temporelle.

5. Procédé selon la revendication 4, **caractérisé en ce que** la normalisation temporelle du résidu comprend la division du résidu par l'estimation de l'enveloppe temporelle.

6. Procédé selon la revendication 4 ou 5, **caractérisé en ce que** la détermination d'une nouvelle enveloppe temporelle pour le résidu comprend une modification (32) de paramètres de l'enveloppe temporelle du résidu selon lesdites consignes de modification et une application (34) de l'enveloppe temporelle modifiée au résidu normalisé.

7. Procédé selon l'une quelconque des revendications 1 et 5, **caractérisé en ce que** l'estimation de l'enveloppe temporelle et la détermination d'une nouvelle enveloppe temporelle sont confondues.

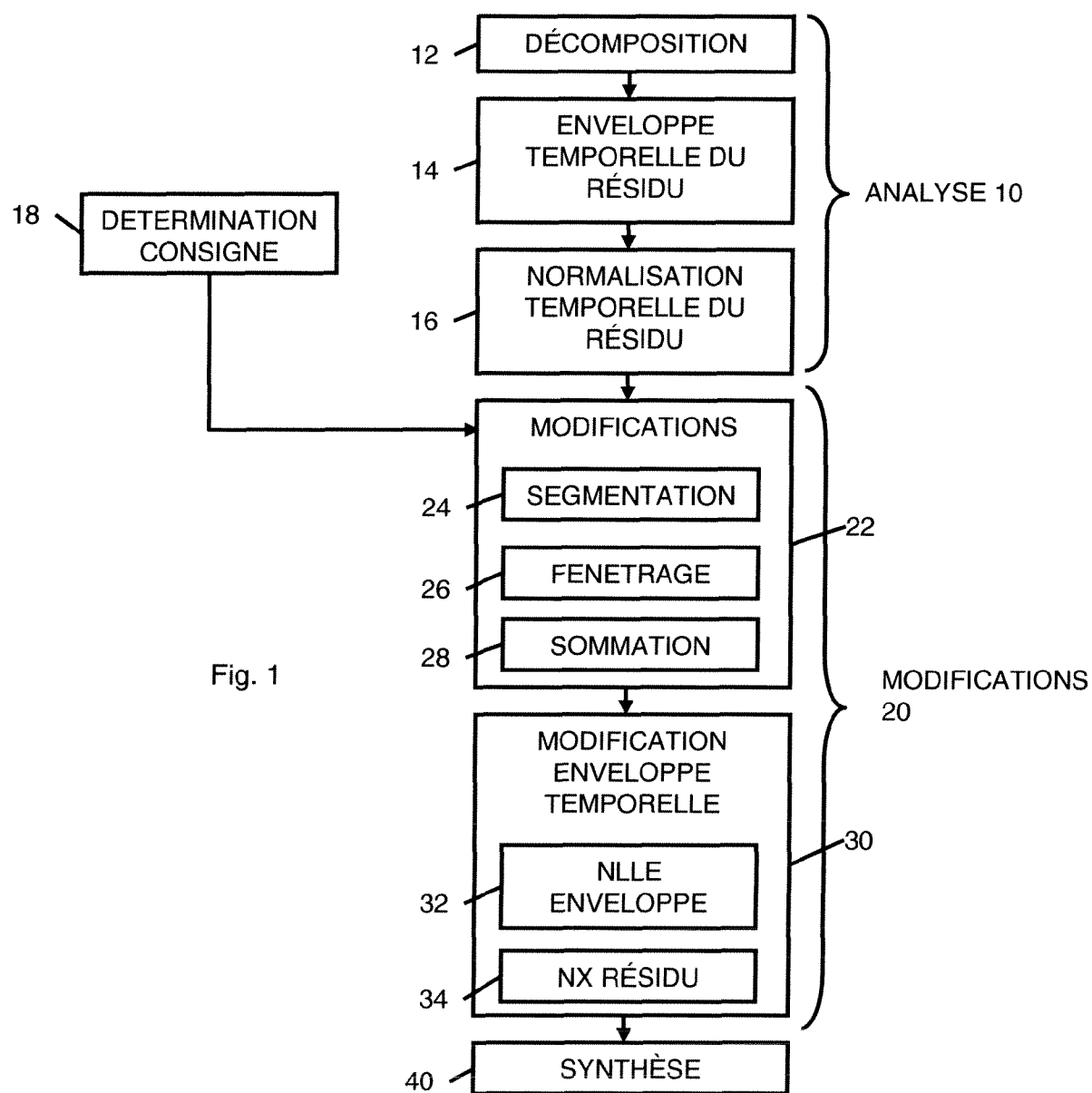
8. Procédé selon l'une quelconque des revendications 1 à 7, **caractérisé en ce que** la modification de caractéristiques acoustiques comprend une modification d'informations de fréquence fondamentale et de durée de la partie paramétrique et du résidu.

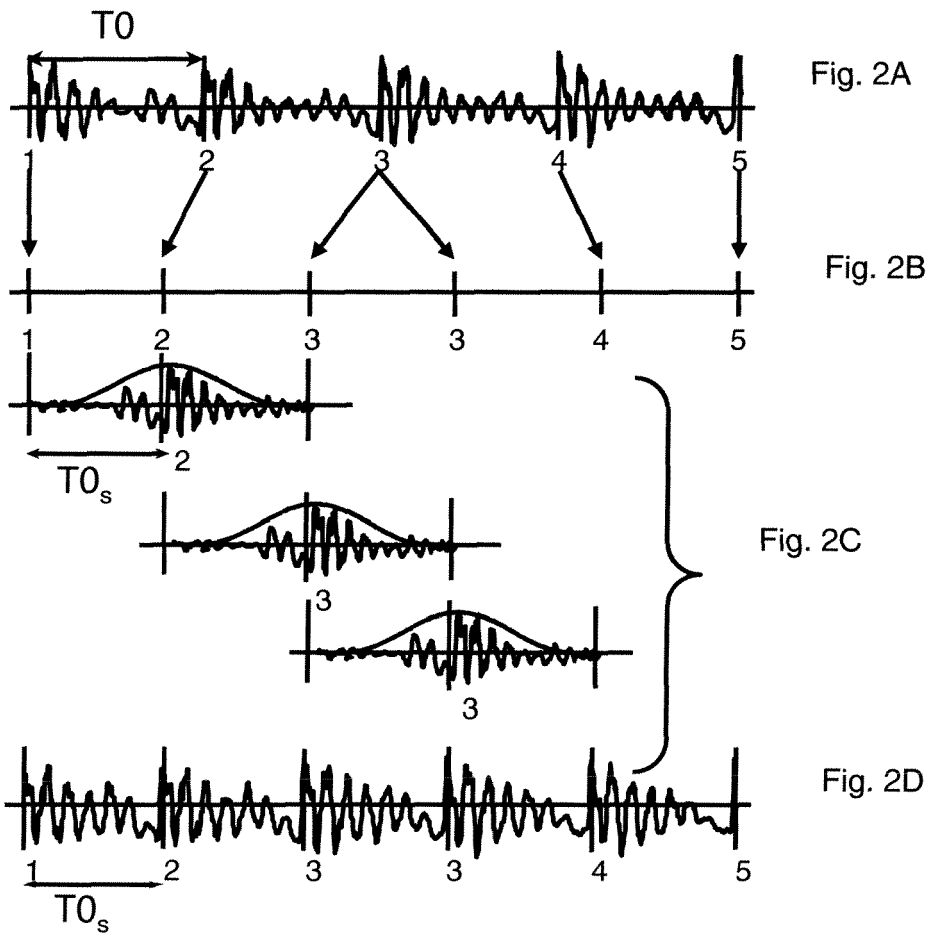
9. Programme pour un dispositif de modification d'un signal de parole ($s(n)$), **caractérisé en ce qu'il** comporte des instructions qui, lorsqu'elles sont exécutées sur un calculateur de ce dispositif, entraînent la mise en oeuvre d'un procédé selon l'une quelconque des revendications 1 à 8.

10. Dispositif de modification d'un signal de parole, **caractérisé en ce qu'il** comporte :

- des moyens de décomposition du signal en une partie paramétrique et un résidu non paramétrique ($e(n)$) ;
- des moyens d'estimation de l'enveloppe temporelle du résidu ;
- des moyens de modification de caractéristiques acoustiques de la partie paramétrique et du résidu selon des consignes de modification ;
- des moyens de détermination, selon lesdites consignes de modification, d'une nouvelle enveloppe temporelle pour le résidu modifié ; et
- des moyens de synthèse d'un signal de parole modifié à partir de la partie paramétrique modifiée et du résidu tel que modifié et avec la nouvelle enveloppe temporelle.

11. Dispositif selon la revendication 10, **caractérisé en ce qu'il** comporte des moyens aptes à la mise en oeuvre d'un procédé selon l'une quelconque des revendications 2 à 8.







Office européen
des brevets

RAPPORT DE RECHERCHE EUROPEENNE

Numéro de la demande
EP 08 10 0453

DOCUMENTS CONSIDERES COMME PERTINENTS			
Catégorie	Citation du document avec indication, en cas de besoin, des parties pertinentes	Revendication concernée	CLASSEMENT DE LA DEMANDE (IPC)
P,X	DAMIEN VINCENT: "Analyse et contrôle du signal glottique en synthèse de la parole." 15 janvier 2007 (2007-01-15), ENST BRETAGNE ET UNIVERSITÉ DE RENNES 1, RENNES, FRANCE, XP002442389 * chapitre 5.3 Modification par modèle ARX-LF *	1-10	INV. G10L13/06 G10L19/08 G10L21/04 ADD. G10L21/00
P,X	VINCENT D ET AL: "A new method for speech synthesis and transformation based on an ARX-LF source-filter decomposition and HNM modeling" 2007 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, 15 avril 2007 (2007-04-15), - 20 avril 2007 (2007-04-20) pages IV-525, XP002442991 HONOLULU, HI, USA ISBN: 1-4244-0727-3 * le document en entier *	1-10	
D,A	YANNIS STYLIANOU: "Applying the Harmonic Plus Noise Model in Concatenative Speech Synthesis" IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, IEEE SERVICE CENTER, NEW YORK, NY, US, vol. 9, no. 1, janvier 2001 (2001-01), XP011054068 ISSN: 1063-6676 * le document en entier *	1-10	DOMAINES TECHNIQUES RECHERCHES (IPC) G10L
A	WO 2006/106466 A (KONINKL PHILIPS ELECTRONICS NV [NL]; HARMA AKI S [NL]) 12 octobre 2006 (2006-10-12) * abrégé * * alinéas [0020], [0040], [0056] *	1-10	
Le présent rapport a été établi pour toutes les revendications			
Lieu de la recherche La Haye		Date d'achèvement de la recherche 27 février 2008	Examineur Quélavoine, Régis
CATEGORIE DES DOCUMENTS CITES X : particulièrement pertinent à lui seul Y : particulièrement pertinent en combinaison avec un autre document de la même catégorie A : arrière-plan technologique O : divulgation non-écrite P : document intercalaire		T : théorie ou principe à la base de l'invention E : document de brevet antérieur, mais publié à la date de dépôt ou après cette date D : cité dans la demande L : cité pour d'autres raisons & : membre de la même famille, document correspondant	

4
EPO FORM 1503 03.82 (P04C02)



DOCUMENTS CONSIDERES COMME PERTINENTS			
Catégorie	Citation du document avec indication, en cas de besoin, des parties pertinentes	Revendication concernée	CLASSEMENT DE LA DEMANDE (IPC)
A	WO 96/16533 A (FINK FLEMING K [DK]; HARTMANN UWE [DK]; HERMANSEN KJELD [DK]; RUBAK PE) 6 juin 1996 (1996-06-06) * abrégé * * page 1, ligne 6-13 * * page 14, ligne 22-31 * -----	1-10	
			DOMAINES TECHNIQUES RECHERCHES (IPC)
Le présent rapport a été établi pour toutes les revendications			
Lieu de la recherche La Haye		Date d'achèvement de la recherche 27 février 2008	Examineur Quélavoine, Régis
CATEGORIE DES DOCUMENTS CITES X : particulièrement pertinent à lui seul Y : particulièrement pertinent en combinaison avec un autre document de la même catégorie A : arrière-plan technologique O : divulgation non-écrite P : document intercalaire T : théorie ou principe à la base de l'invention E : document de brevet antérieur, mais publié à la date de dépôt ou après cette date D : cité dans la demande L : cité pour d'autres raisons & : membre de la même famille, document correspondant			

4

EPO FORM 1503 03.82 (P4C02)

**ANNEXE AU RAPPORT DE RECHERCHE EUROPEENNE
RELATIF A LA DEMANDE DE BREVET EUROPEEN NO.**

EP 08 10 0453

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

27-02-2008

Document brevet cité au rapport de recherche	Date de publication	Membre(s) de la famille de brevet(s)	Date de publication
WO 2006106466 A	12-10-2006	AUCUN	

WO 9616533 A	06-06-1996	AT 179827 T	15-05-1999
		AU 3978595 A	19-06-1996
		DE 69509555 D1	10-06-1999
		DE 69509555 T2	02-09-1999
		DK 0796489 T3	01-11-1999
		EP 0796489 A2	24-09-1997
		JP 10509256 T	08-09-1998
		US 5933801 A	03-08-1999

EPO FORM P0460

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

RÉFÉRENCES CITÉES DANS LA DESCRIPTION

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

Littérature non-brevet citée dans la description

- **Y. STYLIANOU.** Applying the Harmonic plus Noise Model in concatenative speech synthesis. *IEEE Transactions on Speech and Audio Processing*, Janvier 2001, vol. 9 (1), 21-29 [0006]
- **H. L. LU.** Toward a high-quality singing synthesizer with vocal texture control. Stanford University, 2002 [0007]
- **T. F. QUATIERI ; R. B. DUNN ; T. E. HANNA.** Time-scale modification of complex acoustic signals. *ICASSP*, 1993, vol. 1, 213-216 [0008]
- **J. MAKHOUL.** Linear Prediction: a tutorial review. *Proceedings of the IEEE*, Avril 1975, vol. 63 (4), 561-580 [0023]
- **J. DURBIN.** The fitting of time-series models. *Rev. Inst. Int. Statistics* [0031]
- **LILJENCRAANTS ; FANT.** A four-parameter model of glottal flow. *STL-QPSR*, 1985, vol. 4, 1-13 [0033]
- **D. VINCENT ; O. ROSEC ; T. CHONAVEL.** Estimation of LF glottal source parameters based on ARX model. *Interspeech'05*, 2005, 333-336 [0035]
- **E. MOULINES ; J. LAROCHE.** Non-parametric techniques for pitch-scale and time-scale modification of speech. *Speech Communication*, 1995, vol. 16, 175-205 [0052] [0059]
- **D. VINCENT ; O. ROSEC ; T. CHONAVEL.** Glottal closure instant estimation using an appropriateness measure of the source and continuity constraints. *IEEE ICASSP'06*, Mai 2006, vol. 1, 381-384 [0056]