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

(11) **EP 1 526 508 B1**

(12) **FASCICULE DE BREVET EUROPEEN**

(45) Date de publication et mention de la délivrance du brevet:
27.05.2009 Bulletin 2009/22

(51) Int Cl.:
G10L 19/00 *(2006.01)* ***G10L 13/06*** *(2006.01)*

(21) Numéro de dépôt: **04105204.4**

(22) Date de dépôt: **21.10.2004**

(54) **Procédé de sélection d'unités de synthèse**

Verfahren zum Auswählen von Syntheseneinheiten

Method for the selection of synthesis units

(84) Etats contractants désignés:
AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HU IE IT LI LU MC NL PL PT RO SE SI SK TR

(30) Priorité: **24.10.2003 FR 0312494**

(43) Date de publication de la demande:
27.04.2005 Bulletin 2005/17

(73) Titulaire: **Thales**
92200 Neuilly Sur Seine (FR)

(72) Inventeurs:
- **CAPMAN, François,**
  THALES Intellectual Property
  94117 CX, ARCUEIL (FR)
- **PADELLINI, Marc,**
  THALES Intellectual Property
  94117 CX, ARCUEIL (FR)

(74) Mandataire: **Dudouit, Isabelle**
Marks & Clerk France
Conseils en Propriété Industrielle
Immeuble " Visium "
22, avenue Aristide Briand
94117 Arcueil Cedex (FR)

(56) Documents cités:
US-A1- 2002 065 655

- **M. PADELLINI, G. BAUDOIN AND F. CAPMAN: "Codage de la parole a très bas débit par indexation d'unités de taille variable" RJC'2003, 5ÈMES RENCONTRES JEUNES CHERCHEURS EN PAROLE, 23 septembre 2003 (2003-09-23), - 26 septembre 2003 (2003-09-26) XP002285303 GRENOBLE, FRANCE**
- **BAUDOIN G ; EL CHAMI F: "Corpus based very low bit rate speech coding" 2003 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, vol. 1, 6 avril 2003 (2003-04-06), - 10 avril 2003 (2003-04-10) pages 792-795, XP002285304 HONG KONG, CHINA ISBN: 0-7803-7663-3**
- **M. PADELLINI, F. CAPMAN, G. BAUDOIN: "Dynamic Unit Selection for Very Low Bit Rate Coding at 500 bits/sec" TSD 2004, 8 septembre 2004 (2004-09-08), - 11 septembre 2004 (2004-09-11) XP002312723 BRNO, CZ**

EP 1 526 508 B1

Description

[0001] L'invention concerne un procédé de sélection d'unités de synthèse.

[0002] Elle concerne par exemple un procédé de sélection et de codage d'unités de synthèse pour un codeur de parole très bas débit, par exemple inférieur à 600 bits/sec.

[0003] Les techniques d'indexation d'unités de parole naturelle ont récemment permis le développement de systèmes de synthèse à partir du texte particulièrement performants. Ces techniques sont dorénavant étudiées dans le cadre du codage à très bas débit de la parole, conjointement avec des algorithmes empruntés au domaine de la reconnaissance vocale, Ref [1-5]. L'idée principale consiste à identifier dans le signal de parole à coder, une segmentation quasi optimale en unités élémentaires. Ces unités peuvent être des unités obtenues à partir d'une transcription phonétique, qui a l'inconvénient de devoir être corrigée manuellement pour un résultat optimal, ou de façon automatique selon des critères de stabilité spectrale. A partir de ce type de segmentation, et pour chacun des segments, on cherche l'unité de synthèse la plus proche dans un dictionnaire obtenu lors d'une phase d'apprentissage préalable, et contenant des unités de synthèse de référence.

[0004] Le schéma de codage utilisé consiste à modéliser l'espace acoustique du locuteur (ou des locuteurs) par des modèles de Markov cachés (HMM ou Hidden Markov Models). Ces modèles dépendants ou indépendants du locuteur sont obtenus lors d'une phase d'apprentissage préalable à partir d'algorithmes identiques à ceux mis en oeuvre dans les systèmes de reconnaissance de la parole. La différence essentielle réside dans le fait que les modèles sont appris sur des vecteurs regroupés par classes de façon automatique et non de manière supervisée à partir d'une transcription phonétique. La procédure d'apprentissage consiste alors à obtenir de façon automatique la segmentation des signaux d'apprentissage (par exemple en utilisant la méthode dite de décomposition temporelle), à regrouper les segments obtenus dans un nombre fini de classes correspondant au nombre de modèles HMM que l'on souhaite construire. Le nombre de modèles est directement lié à la résolution recherchée pour représenter l'espace acoustique du ou des locuteurs. Une fois obtenus, ces modèles permettent de segmenter le signal à coder en utilisant un algorithme de Viterbi. La segmentation permet d'associer à chaque segment, l'indice de classe et sa longueur. Cette information n'étant pas suffisante pour modéliser l'information spectrale, pour chacune des classes une réalisation de trajectoire spectrale est sélectionnée parmi plusieurs unités dites de synthèse. Ces unités sont extraites de la base d'apprentissage lors de sa segmentation utilisant les modèles HMM. Il est possible de tenir compte du contexte par exemple en utilisant plusieurs sous-classes permettant de prendre en compte les transitions d'une classe vers l'autre. Un premier indice indique la classe à laquelle appartient le segment considéré, un deuxième indice précise la sous-classe à laquelle il appartient comme étant l'indice de classe du segment précédent. L'indice de sous-classe n'est donc pas à transmettre, et l'indice de classe doit être mémorisé pour le segment suivant. Les sous-classes ainsi définies permettent de tenir compte des différentes transitions vers la classe associée au segment considéré. A l'information spectrale on ajoute l'information de prosodie, c'est-à-dire la valeur des paramètres de pitch et d'énergie et leurs évolutions.

[0005] Dans l'optique de réaliser un codeur très bas débit, il est nécessaire d'optimiser l'allocation des bits et donc du débit entre les paramètres associés à l'enveloppe spectrale et à l'information de prosodie. Une méthode classique consiste dans un premier temps à sélectionner l'unité la plus proche d'un point de vue spectral puis, une fois l'unité sélectionnée, à coder l'information de prosodie, soit de façon indépendante de l'unité sélectionnée.

[0006] Le document intitulé « codage de la parole à très bas débit par indexation d'unités de taille variable » septembre 2003, rencontre des jeunes chercheurs en parole décrit un procédé dans lequel l'idée principale consiste à segmenter le signal de parole aux frontières de ce segment pour constituer des unités de synthèse. Une fois classées et regroupées dans une base de données, le signal de parole peut être caractérisé par une suite d'index faisant référence aux unités de la base. La base de données correspond aux données d'apprentissage. Afin de caractériser les variations d'intonation, la prosodie est extraite du signal à coder et est transmise avec les index

[0007] Le document « Codage de la parole à très bas débit par indexation d'unités de taille variable" C.Baverel, P.Gournay, F.Capman, G.Chollet IST OTAN, 2001 concerne une démonstration de concept de la concaténation d'unités.

[0008] Le procédé selon la présente invention propose une nouvelle méthode de sélection de l'unité de synthèse la plus proche conjointement à la modélisation et à la quantification des informations supplémentaires nécessaires au niveau du décodeur pour la restitution du signal de parole.

[0009] L'invention concerne un procédé de sélection d'unités de synthèse d'une information d'une information se présentant sous la forme d'un segment de parole à coder et pouvant être décomposée en unités de synthèse. Il comporte au moins les étapes suivantes :

pour un segment d'information considéré :

- déterminer la valeur F0 de la fréquence fondamentale moyenne pour le segment d'information considéré,
- sélectionner un sous-ensemble d'unités de synthèse défini comme étant celui dont les valeurs moyennes de pitch sont les plus proches de la valeur de pitch F0,

- appliquer un ou plusieurs critères de proximité aux unités de synthèse sélectionnées pour déterminer une unité de synthèse représentative du segment d'information.

[0010] Selon une variante on utilise comme critères de proximité la fréquence fondamentale ou pitch, ou la distorsion spectrale, et/ou le profil d'énergie et on exécute une étape de fusion des critères utilisés afin de déterminer l'unité de synthèse représentative.

[0011] Le procédé comporte par exemple une étape de codage et/ou une étape de correction du pitch par modification du profil de synthèse.

[0012] L'étape de codage et/ou correction du pitch peut être une transformation linéaire du profil du pitch d'origine.

[0013] Le procédé est par exemple utilisé pour la sélection et/ou le codage d'unités de synthèse pour un codeur de parole très bas débit.

[0014] L'invention présente notamment les avantages suivants :

- le procédé permet d'optimiser le débit alloué à l'information de prosodie dans le domaine de la parole.
- il permet de conserver, lors de la phase de codage, l'intégralité des unités de synthèse déterminées lors de la phase d'apprentissage avec cependant un nombre de bits constant pour coder l'unité de synthèse.
- Dans un schéma de codage indépendant du locuteur, ce procédé offre la possibilité de couvrir l'ensemble des valeurs de pitch possibles (ou fréquences fondamentales) et de sélectionner l'unité de synthèse en tenant compte en partie des caractéristiques du locuteur.
- La sélection peut s'appliquer à tout système basé sur une sélection d'unités et donc aussi à un système de synthèse à partir du texte.

[0015] D'autres caractéristiques et avantages de l'invention apparaîtront mieux à la lecture de la description qui suit d'un exemple de réalisation non limitatif annexé des figures qui représentent :

- La figure 1 un schéma de principe de sélection de l'unité de synthèse associée au segment d'information à coder,
- La figure 2 un schéma de principe d'estimation des critères de similarité pour le profil du pitch,
- La figure 3 un schéma de principe d'estimation des critères de similarité pour le profil énergétique,
- La figure 4 un schéma de principe d'estimation des critères de similarité pour l'enveloppe spectrale,
- La figure 5 un schéma de principe du codage du pitch par correction du profil de pitch de synthèse.

[0016] Afin de mieux faire comprendre l'idée mise en oeuvre dans la présente l'invention, l'exemple qui suit est donné à titre illustratif et nullement limitatif pour un procédé mis en oeuvre dans un vocodeur, en particulier la sélection et le codage d'unités de synthèse pour un codeur de parole très bas débit.

[0017] Pour rappel, au niveau d'un vocodeur, le signal de parole est analysé trame à trame afin d'extraire les paramètres caractéristiques (paramètres spectraux, pitch, énergie). Cette analyse se fait classiquement à l'aide d'une fenêtre glissante définie sur l'horizon de la trame. Cette trame a une durée de l'ordre de 20 ms, et la mise à jour se fait avec un décalage de la fenêtre d'analyse de l'ordre de 10ms à 20 ms.

[0018] Lors d'une phase d'apprentissage, un ensemble de modèles de Markov cachés (HMM, Hidden Markov Model) sont appris. Ils permettent de modéliser des segments de parole (ensemble de trames successives) pouvant être associés à des phonèmes si la phase d'apprentissage est supervisée (segmentation et transcription phonétique disponibles) ou à des sons spectralement stables dans le cas d'une segmentation obtenue de façon automatique. On utilise ici 64 modèles HMM, qui permettent lors de la phase de reconnaissance d'associer à chaque segment l'indice du modèle HMM identifié, et donc la classe à laquelle il appartient. Les modèles HMM servent aussi à l'aide d'un algorithme de type Viterbi à réaliser lors de la phase de codage la segmentation et la classification de chacun des segments (appartenance à une classe). Chaque segment est donc identifié par un indice compris entre 1 et 64 qui est transmis au décodeur.

[0019] Le décodeur utilise cet indice pour retrouver l'unité de synthèse dans le dictionnaire construit lors de la phase d'apprentissage. Les unités de synthèse qui constituent le dictionnaire sont simplement les séquences de paramètres associés aux segments obtenus sur le corpus d'apprentissage.

[0020] Une classe du dictionnaire contient l'ensemble des unités associées à un même modèle HMM. Chaque unité de synthèse est donc caractérisée par une séquence de paramètres spectraux, une séquence de valeur de pitch (profil de pitch), une séquence de gains (profil énergétique).

[0021] Afin d'améliorer la qualité de la synthèse, chaque classe (de 1 à 64) du dictionnaire est subdivisée en 64 sous-classes, où chaque sous-classe contient les unités de synthèse qui sont précédées temporellement par un segment appartenant à une même classe. Cette approche permet de tenir compte du contexte passé, et donc d'améliorer la restitution des zones transitoires d'une unité vers l'autre.

[0022] La présente invention concerne notamment un procédé de sélection d'une unité de synthèse multicritères. Le procédé permet par exemple de tenir compte simultanément du pitch, de la distorsion spectrale, et des profils d'évolution du pitch et de l'énergie.

[0023] Le procédé de sélection pour un segment de parole à coder comporte par exemple les étapes de sélection schématisées à la figure 1 :

1) Extraire le pitch moyen F_0 (fréquence fondamentale moyenne) sur le segment à coder composé de plusieurs trames. Le pitch est par exemple calculé pour chaque trame T, les erreurs de pitch sont corrigées en tenant compte de l'ensemble du segment afin d'éliminer les erreurs de détection voisé/non voisé, et le pitch moyen est calculé sur l'ensemble des trames voisées du segment.
Il est possible de représenter le pitch sur 5 bits, en utilisant par exemple un quantificateur non uniforme (compression logarithmique) appliqué à la période de pitch.
La valeur du pitch de référence est par exemple obtenue à partir d'un générateur de prosodie dans le cas d'une application en synthèse.
2) la valeur de pitch moyen F_0 étant ainsi quantifiée, sélectionner un sous-ensemble d'unités de synthèse SE dans la sous-classe considérée. Le sous-ensemble est défini comme étant celui dont les valeurs moyennes de pitch sont les plus proches de la valeur de pitch F_0.
Dans la configuration précédente cela conduit à retenir de façon systématique les 32 unités les plus proches selon le critère du pitch moyen. Il est donc possible de retrouver ces unités au niveau du décodeur à partir du pitch moyen transmis.
3) Parmi les unités de synthèse ainsi sélectionnées, appliquer un ou plusieurs critères de proximité ou de similarité, par exemple le critère de distorsion spectrale, et/ou le critère de profil d'énergie et/ou le critère de pitch pour déterminer l'unité de synthèse.
Lorsque l'on utilise plusieurs critères, une étape de fusion 3b) est réalisée pour prendre la décision. L'étape de fusion des différents critères est réalisée par combinaison linéaire ou non-linéaire. Les paramètres utilisés pour réaliser cette combinaison peuvent être obtenus par exemple sur un corpus d'apprentissage en minimisant un critère de distorsion spectrale sur le signal re-synthétisé. Ce critère de distorsion peut avantageusement inclure une pondération perceptuelle soit au niveau des paramètres spectraux utilisés, soit au niveau de la mesure de distorsion. Dans le cas d'une loi de pondération non linéaire il est possible d'utiliser un réseau connexionniste (MLP, Multi Layer Perceptron par exemple), de la logique floue, ou une autre technique.
4) Etape de codage du pitch

[0024] Le procédé peut comporter dans une variante de réalisation une étape de codage de pitch par correction du profil de pitch de synthèse exposée en détail ci-après.

[0025] Le critère relatif au profil d'évolution du pitch permet en partie de tenir compte de l'information de voisement. Il est cependant possible de le désactiver lorsque le segment est totalement non voisé, ou que la sous-classe sélectionnée est aussi non voisée. En effet, on peut remarquer principalement trois types de sous-classes : les sous-classes contenant majoritairement des unités voisées, celles contenant majoritairement des unités non voisées, et les sous-classes contenant majoritairement des unités mixtes.

[0026] Le procédé selon l'invention ne se limite pas à optimiser le débit alloué à l'information de prosodie mais permet aussi de conserver pour la phase de codage l'intégralité des unités de synthèse obtenues lors de la phase d'apprentissage avec un nombre de bits constant pour coder l'unité de synthèse. En effet l'unité de synthèse est caractérisée à la fois par la valeur de pitch et par son indice. Cette approche permet dans un schéma de codage indépendant du locuteur de couvrir l'ensemble des valeurs de pitch possibles et de sélectionner l'unité de synthèse en tenant compte en partie des caractéristiques du locuteur, il existe en effet pour un même locuteur une corrélation entre la plage de variation du pitch et les caractéristiques du conduit vocal (en particulier la longueur).

[0027] On peut remarquer que le principe de sélection d'unités décrit peut s'appliquer à tout système dont le fonctionnement est basé sur une sélection d'unités et donc aussi à un système de synthèse à partir du texte.

[0028] La figure 2 schématise un principe d'estimation des critères de similarité pour le profil du pitch.
Le procédé comporte par exemple les étapes suivantes :

A1) sélectionner dans la sous-classe identifiée du dictionnaire des unités de synthèse et à partir de la valeur moyenne du pitch, les N unités les plus proches au sens du critère du pitch moyen. La suite du traitement se fait alors sur les profils de pitch associés à ces N unités. Le pitch est extrait lors de la phase d'apprentissage sur les unités de synthèse, et lors de la phase de codage sur le signal à coder. Les méthodes possibles pour l'extraction du pitch sont nombreuses, cependant les méthodes hybrides, combinant un critère temporel (AMDF, Average Magnitude Difference Function, ou autocorrélation normalisée) et un critère fréquentiel (HPS, Harmonic Power Sum, structure en peigne, ...) sont potentiellement plus robustes.
A2) aligner temporellement les N profils avce celui du segment à coder, par exemple par interpolation linéaire des N profils. Il est possible d'utiliser une technique d'alignement plus optimale basée sur un algorithme de programmation dynamique (DTW ou Dynamic Time Warping). L'algorithme s'applique sur les paramètres spectraux, les autres paramètres pitch, énergie, etc sont alignés de manière synchrone aux paramètres spectraux. Dans ce cas il faut transmettre les informations relatives au chemin d'alignement.
A3) calculer N mesures de similarités, entre les N profils de pitch alignés et le profil de pitch du segment de parole à coder pour obtenir les N coefficients de similarité {rp(1), rp(2), ...rp(N)}. Cette étape peut être réalisée au moyen d'une intercorrélation norma-

lisée.

L'alignement temporel peut être un alignement par ajustement simple des longueurs (interpolation linéaire des paramètres). L'utilisation d'une simple correction des longueurs des unités de synthèse permet notamment de ne pas transmettre d'information relative au chemin d'alignement, le chemin d'alignement étant partiellement pris en compte par les corrélations des profils de pitch et d'énergie.

Dans le cas de segments mixtes (co-existence au sein d'un même segment de trames voisées et non voisées), l'utilisation des trames non voisées pour lesquelles le pitch est arbitrairement positionné à zéro permet de tenir compte dans une certaine mesure de l'évolution du voisement.

La figure 3 schématise le principe d'estimation des critères de similarité pour le profil énergétique.

Le procédé comporte par exemple les étapes suivantes :

A4) extraire les profils d'évolution de l'énergie pour les N unités sélectionnées comme indiqué précédemment, c'est-à-dire selon un critère de proximité du pitch moyen. Selon la technique de synthèse utilisée, le paramètre d'énergie utilisé peut soit correspondre à un gain (associé à un filtre de type LPC par exemple) ou une énergie (l'énergie calculée sur la structure harmonique dans le cas d'une modélisation harmonique/stochastique du signal). Enfin, l'estimation de l'énergie peut avantageusement se faire de manière synchrone du pitch (1 valeur d'énergie par période de pitch). Les profils énergétiques sont pré-calculés pour les unités de synthèse lors de la phase d'apprentissage.

A5) aligner temporellement les N profils avec celui des segments à coder, par exemple par interpolation linéaire, ou par programmation dynamique (alignement non-linéaire) de façon similaire à la méthode mise en oeuvre pour corriger le pitch.

A6) calculer N mesures de similarités, entre les profils des N valeurs d'énergie alignées et le profil d'énergie du segment de parole à coder pour obtenir les N coefficients de similarité {re(1), re(2), ..., re (N)}. Cette étape peut aussi être réalisée au moyen d'une intercorrélation normalisée.

La figure 4 schématise le principe d'estimation des critères de similarité pour l'enveloppe spectrale.

Le procédé comporte les étapes suivantes :

A7) aligner temporellement les N profils,

A8) déterminer les profils d'évolution des paramètres spectraux pour les N unités sélectionnées comme indiqué précédemment, c'est-à-dire selon un critère de proximité du pitch moyen. Il s'agit ici tout simplement de calculer le pitch moyen du segment à coder, et de considérer les unités de synthèse de la sous-classe associée (indice HMM courant pour définir la classe, indice HMM précédent pour définir la sous-classe) qui ont un pitch moyen proche.

A9) calculer N mesures de similarités, entre la séquence spectrale du segment à coder et les N séquences spectrales extraites des unités de synthèse sélectionnées pour obtenir les N coefficients de similarité {rs(1), rs(2), ..., rs(N)}. Cette étape peut être réalisée au moyen d'une intercorrélation normalisée.

[0029] La mesure de similarité peut être une distance spectrale.

[0030] L'étape A9) comprend par exemple une étape où l'on moyenne l'ensemble des spectres d'un même segment et la mesure de similarité est une mesure d'intercorrélation.

[0031] Le critère de distorsion spectrale est par exemple calculé sur des structures harmoniques ré-échantillonnées à pitch constant ou ré-échantillonnées au pitch du segment à coder, après interpolation des structures harmoniques initiales.

[0032] Le critère de similarité va dépendre des paramètres spectraux utilisés (par exemple du type de paramètres utilisés pour la représentation de l'enveloppe). Plusieurs types de paramètres spectraux peuvent être utilisés, dans la mesure où ils permettent de définir une mesure de distorsion spectrale. Dans le domaine du codage de la parole, il est courant d'utiliser les paramètres LSP ou LSF (LSP, Line Spectral Pair, LSF, Line Spectral Frequencies) dérivés d'une analyse par prédiction linéaire. Dans le domaine de la reconnaissance vocale, les paramètres cepstraux sont généralement utilisés, et ils peuvent soit être dérivés d'une analyse par prédiction linéaire (LPCC, Linear Prediction Cepstrum Coefficients) ou estimés à partir d'un banc de filtres souvent sur une échelle perceptuelle de type Mel ou Bark (MFCC, Mel Frequency Cepstrum Coefficients). Il est aussi possible dans la mesure où on utilise une modélisation sinusoïdale de la composante harmonique du signal de parole, d'utiliser directement les amplitudes des fréquences harmoniques. Ces derniers paramètres étant estimés en fonction du pitch ne peuvent être utilisés directement pour calculer une distance. Le nombre de coefficients obtenus est en effet variable en fonction du pitch, contrairement aux paramètres LPCC, MFCC ou LSF. Un prétraitement consiste alors à estimer une enveloppe spectrale à partir des amplitudes harmoniques (interpolation linéaire ou polynomiale de type spline) et à réechantillonner l'enveloppe ainsi obtenue, soit en utilisant la fréquence fondamentale du segment à coder, soit en utilisant une fréquence fondamentale constante (100 Hz par exemple). Une fréquence fondamentale constante permet de pré-calculer l'ensemble des structures harmoniques des unités de synthèse lors de la phase d'apprentissage. Le re-échantillonnage se fait alors uniquement sur le segment à coder. D'autre part, si on se limite à un alignement temporel par interpolation linéaire, il est possible de moyenner les structures harmoniques sur l'ensemble des segments considérés. La mesure de similarité peut alors être estimée simplement à partir de la structure harmonique moyenne du segment à coder, et celle de l'unité de synthèse considérée. Cette mesure de

similarité peut aussi être une mesure d'intercorrélation normalisée. On peut aussi noter que la procédure de réechantillonnage peut s'effectuer sur une échelle perceptuelle des fréquences (Mel ou Bark).

[0033] Pour la procédure d'alignement temporel il est possible d'utiliser soit un algorithme de programmation dynamique (DTW, Dynamic Time Warping), soit d'effectuer une interpolation linéaire simple (ajustement linéaire des longueurs). Dans l'hypothèse où l'on ne souhaite pas transmettre d'information supplémentaire relative au chemin d'alignement, il est préférable d'utiliser une simple interpolation linéaire des paramètres. La prise en compte du meilleur alignement est alors en partie réaliser par la procédure de sélection.

Codage du pitch par modification du profil de synthèse

[0034] Selon un mode de réalisation, le procédé comporte une étape de codage du pitch par modification du profil de synthèse. Cela consiste à resynthétiser un profil de pitch à partir de celui de l'unité de synthèse sélectionnée et un gain linéairement variable sur la durée du segment à coder. Il suffit alors de transmettre une valeur supplémentaire pour caractériser le gain correcteur sur l'ensemble du segment.

[0035] Le pitch reconstruit au niveau du décodeur est donné par l'équation suivante :

$$\hat{f}_0(n) = g(n).f_{0S}(n) = (a.n + b).f_{0S}(n) \quad (1)$$

où $f_{0S}(n)$ est le pitch à la trame d'indice n de l'unité de synthèse.

Cela correspond à une transformation linéaire du profil du pitch.

Les valeurs optimales de a et b sont estimées au niveau du codeur en minimisant l'erreur quadratique moyenne :

$$\sum_n e_0^2(n) = \sum_n \left[f_0(n) - \hat{f}_0(n)\right]^2 \quad (2)$$

ce qui conduit aux relations suivantes :

$$a = \frac{(S_4.S_2 - S_5.S_1)}{(S_2.S_2 - S_3.S_1)} \quad (3)$$

et

$$b = \frac{(S_5.S_2 - S_4.S_3)}{(S_2.S_2 - S_3.S_1)} \quad (4)$$

où

$$S_1 = \sum_n f_{0S}(n).f_{0S}(n)$$

$$S_2 = \sum_n n.f_{0S}(n).f_{0S}(n)$$

$$S_3 = \sum_n n^2.f_{0S}(n).f_{0S}(n)$$

$$S_4 = \sum_n f_0(n).f_{0S}(n)$$

$$S_5 = \sum_n n.f_0(n).f_{0S}(n)$$

Le coefficient a, ainsi que la valeur moyenne du pitch modélisé sont quantifiés et transmis :

$$a_q = Q[a] \quad (5)$$

$$f_{0q} = Q\left[\frac{\sum_n (a.n + b).f_{0S}(n)}{N}\right] \quad (6)$$

La valeur du coefficient b est obtenu au niveau du décodeur à partir de la relation suivante :

$$b_q = \frac{f_{0q} - \frac{\sum a_q.n.f_{0S}(n)}{N}}{\langle f_{0S} \rangle} \quad (7)$$

où $\langle f_{0S} \rangle$ est le pitch moyen de l'unité de synthèse.

Remarque : cette méthode de correction peut bien entendu s'appliquer au profil énergétique.

Exemple de débit associé au schéma de codage

[0036] Les informations relatives au débit associé au schéma de codage décrit précédemment sont les suivantes :

- Indice de classe sur 6 bits (64 classes)

- Indice de l'unité sélectionnée sur 5 bits (32 unités par sous-classe)
- Longueur du segment sur 4 bits (de 3 à 18 trames)

[0037] Le nombre moyen de segments par seconde se situe entre 15 et 20; ce qui conduit à un débit de base situé entre 225 et 300 bits/sec pour la configuration précédente. A ce débit de base vient s'ajouter le débit nécessaire pour représenter l'information de pitch et d'énergie.

- FO moyen sur 5 bits
- Coefficient correcteur du profil de pitch sur 5bits
- Gain correcteur sur 5 bits

[0038] Le débit associé à la prosodie se situe alors entre 225 et 300 bits/sec, ce qui conduit à un débit global entre 450 et 600 bits/sec.

Références

[0039]

[1] G. Baudoin, F. El Chami, "Corpus based very low bit rate speech coder", Proc. Conf. IEEE ICASSP 2003, Hong-Kong, 2003.

[2] G. Baudoin, J. Cernocky, P. Gournay, G. Chollet, "Codage de la parole à bas et très bas debit", Annales des télécommunications, Vol. 55, N 9-10 Pages 421-456, Nov. 2000.

[3] G. Baudoin, F. Capman, J. Cernocky, F. El-chami, M. Charbit, G. Chollet, D. Petrovska-Delacrétaz. "Advances in Very Low Bit Rate Speech Coding using Recognition and Synthesis Techniques", TSD' 2002, pp. 269-276, Brno, Czech Republic, Sept 2002.

[4] K.Lee, R.Cox, 'A segmental coder based on a concatenative TTS", in Speech Communications, Vol. 38, pp 89-100, 2002.

[5] K.Lee, R.Cox, "A very low bit rate speech coder based on a recognition/synthesis paradigm", in IEEE on ASSP, Vol; 9, pp 482-491, July 2001.

Revendications

1. Procédé de sélection d'unités de synthèse d'une information se présentant sous la forme d'un segment de parole à coder et pouvant être décomposée en unités de synthèse **caractérisé en ce qu'**il comporte au moins les étapes suivantes :

pour un segment d'information considéré :

• déterminer la valeur F0 de la fréquence fondamentale moyenne pour le segment d'information considéré,
• sélectionner un sous-ensemble d'unités de synthèse défini comme étant celui dont les valeurs moyennes de pitch sont les plus proches de la valeur de pitch F0,
• appliquer un ou plusieurs critères de proximité aux unités de synthèse sélectionnées pour déterminer une unité de synthèse représentative du segment d'information.

2. Procédé de sélection d'unités de synthèse selon la revendication 1 **caractérisé en ce que** l'on utilise comme critères de proximité la fréquence fondamentale ou pitch, la distorsion spectrale, et/ou le profil d'énergie et on exécute une étape de fusion des critères afin de déterminer l'unité de synthèse représentative.

3. Procédé de sélection d'unités selon la revendication 1 **caractérisé en ce que** pour un segment de parole à coder le pitch de référence est obtenu à partir d'un générateur de prosodie.

4. Procédé selon la revendication 2 **caractérisé en ce que** l'estimation du critère de similarité pour le profil du pitch comporte au moins les étapes suivantes :

A1) sélectionner dans la sous-classe identifiée du dictionnaire des unités de synthèse et à partir de la valeur moyenne du pitch, les N unités les plus proches au sens du critère du pitch moyen,
A2) aligner temporellement les N profils avec celui du segment à coder,
A3) calculer N mesures de similarités, entre les N profils de pitch alignés et le profil du pitch du segment de parole à coder pour obtenir les N coefficients de similarité {rp(1), rp(2), ... rp(N)}.

5. Procédé selon la revendication 2 **caractérisé en ce que** l'estimation de similarité pour le profil énergétique comporte au moins les étapes suivantes :

A4) déterminer les profils d'évolution de l'énergie pour les N unités sélectionnées selon un critère de proximité du pitch moyen.
A5) aligner temporellement les N profils avec celui du segment à coder,
A6) calculer N mesures de similarités, entre les N profils de d'énergie alignés et le profil d'énergie du segment de parole à coder pour obtenir les N coefficients de similarité {re(1), re(2), ..., re(N)}.

6. Procédé selon la revendication 2 **caractérisé en ce que** l'estimation des critères de similarité pour l'enveloppe spectrale comporte au moins les étapes

suivantes :

A7) aligner temporellement les N profils avec celui du segment à coder,
A8) déterminer les profils d'évolution des paramètres spectraux pour les N unités sélectionnées selon un critère de proximité du pitch moyen,
A9) calculer N mesures des similarités, entre la séquence spectrale du segment à coder et les N séquences spectrales extraites correspondant du segment de parole à coder pour obtenir les N coefficients de similarité {rs(1), rs(2), ..., rs(N)},

7. Procédé selon l'une des revendications 4, 5 et 6 **caractérisé en ce que** l'alignement temporel est un alignement temporel obtenu par programmation dynamique (DTW) ou un alignement par ajustement linéaire des longueurs.

8. Procédé selon l'une des revendications 4, 5 et 6 **caractérisé en ce que** la mesure de similarité est une mesure d'intercorrélation normalisée.

9. Procédé selon la revendication 6 **caractérisé en ce que** la mesure de similarité est une mesure de distance spectrale.

10. Procédé selon la revendication 6 **caractérisé en ce que** l'étape A9) comprend une étape où l'ensemble des spectres d'un même segment est moyenné et **en ce que** la mesure de similarité est une mesure d'intercorrélation.

11. Procédé selon la revendication 6 **caractérisé en ce que** le critère de distorsion spectrale est calculé sur des structures harmoniques ré-échantillonnées à pitch constant ou ré-échantillonnées au pitch du segment à coder, après interpolation des structures harmoniques Initiales.

12. Procédé selon l'une des revendications 1 à 11 **caractérisé en ce qu'**il comporte une étape de codage et/ou une étape de correction du pitch par modification du profil de synthèse.

13. Procédé selon la revendication 12 **caractérisé en ce que** l'étape de codage et/ou correction du pitch est une transformation linéaire du profil du pitch d'origine.

14. Utilisation du procédé selon l'une des revendications 1 à 12 pour la sélection et/ou le codage d'unités de synthèse pour un codeur de parole très bas débit.

Claims

1. Method for selecting synthesis units of an item of information taking the form of a speech segment to be coded and able to be decomposed into synthesis units, **characterized in that** it comprises at least the following steps:

for an information segment considered:

• determining the value F0 of the mean fundamental frequency for the information segment considered,
• selecting a subset of synthesis units defined as being the subset whose mean pitch values are closest to the pitch value F0,
• applying one or more proximity criteria to the selected synthesis units so as to determine a synthesis unit representative of the information segment.

2. Method for selecting synthesis units according to Claim 1, **characterized in that** the fundamental frequency or pitch, the spectral distortion, and/or the energy profile are used as proximity criteria and a step of merging the criteria is executed so as to determine the representative synthesis unit.

3. Method for selecting units according to Claim 1, **characterized in that** for a speech segment to be coded the reference pitch is obtained on the basis of a prosody generator.

4. Method according to Claim 2, **characterized in that** the estimation of the similarity criterion for the profile of the pitch comprises at least the following steps:

A1) selecting from the identified sub-class of the dictionary of the synthesis units and on the basis of the mean value of the pitch, the N closest units within the sense of the mean pitch criterion,
A2) temporally aligning the N profiles with that of the segment to be coded,
A3) calculating N measures of similarity, between the N aligned pitch profiles and the profile of the pitch of the speech segment to be coded so as to obtain the N similarity coefficients {rp (1), rp(2),....rp(N)}.

5. Method according to Claim 2, **characterized in that** the similarity estimation for the energy profile comprises at least the following steps:

A4) determining the evolution profiles of the energy for the N units selected according to a proximity of the mean pitch criterion,
A5) temporally aligning the N profiles with that of the segment to be coded,

A6) calculating N measures of similarity, between the N aligned energy profiles and the energy profile of the speech segment to be coded so as to obtain the N similarity coefficients {re (1), re(2), ..., re(N)}.

6. Method according to Claim 2, **characterized in that** the estimation of the similarity criteria for the spectral envelope comprises at least the following steps:

A7) temporally aligning the N profiles with that of the segment to be coded,
A8) determining the evolution profiles of the spectral parameters for the N units selected according to a proximity of the mean pitch criterion,
A9) calculating N measures of the similarities, between the spectral sequence of the segment to be coded and the corresponding N spectral sequences extracted from the speech segment to be coded so as to obtain the N similarity coefficients {rs(1), rs(2), ..., rs(N)}.

7. Method according to one of Claims 4, 5 and 6, **characterized in that** the temporal alignment is a temporal alignment obtained by dynamic programming (DTW) or an alignment by linear adjustment of the lengths.

8. Method according to one of Claims 4, 5 and 6, **characterized in that** the similarity measure is a normalized intercorrelation measure.

9. Method according to Claim 6, **characterized in that** the similarity measure is a spectral distance measure.

10. Method according to Claim 6, **characterized in that** step A9) comprises a step where the set of the spectra of one and the same segment is averaged and **in that** the similarity measure is an intercorrelation measure.

11. Method according to Claim 6, **characterized in that** the spectral distortion criterion is calculated on harmonic structures re-sampled at constant pitch or re-sampled at the pitch of the segment to be coded, after interpolation of the initial harmonic structures.

12. Method according to one of Claims 1 to 11, **characterized in that** it comprises a step of coding and/or a step of correcting the pitch by modification of the synthesis profile.

13. Method according to Claim 12, **characterized in that** the step of coding and/or correcting the pitch is a linear transformation of the profile of the original pitch.

14. Use of the method according to one of Claims 1 to 12 for the selection and/or the coding of synthesis units for a very low bit rate speech coder.

Patentansprüche

1. Verfahren zur Auswahl von Syntheseeinheiten einer Information, die in Form eines zu codierenden Sprachsegments vorliegt und in Syntheseeinheiten zerlegt werden kann, **dadurch gekennzeichnet, dass** es mindestens die folgenden Schritte aufweist:

für ein betrachtetes Informationssegment:

- Bestimmen des Werts F0 der mittleren Grundfrequenz für das betrachtete Informationssegment,
- Auswählen einer Untergruppe von Syntheseeinheiten, die als diejenige definiert wird, deren mittlere Pitch-Werte dem Pitch-Wert F0 am nächsten sind,
- Anwenden eines oder mehrerer Näherungskriterien an die ausgewählten Syntheseeinheiten, um eine Syntheseeinheit zu bestimmen, die für das Informationssegment repräsentativ ist.

2. Verfahren zur Auswahl von Syntheseeinheiten nach Anspruch 1, **dadurch gekennzeichnet, dass** als Näherungskriterien die Grundfrequenz oder Pitch, die spektrale Verzerrung, und/oder das Energieprofil verwendet werden, und ein Schritt des Mischens der Kriterien ausgeführt wird, um die repräsentative Syntheseeinheit zu bestimmen.

3. Verfahren zur Auswahl von Syntheseeinheiten nach Anspruch 1, **dadurch gekennzeichnet, dass** für ein zu codierendes Sprachsegment der Bezugspitch ausgehend von einem Prosodiegenerator erhalten wird.

4. Verfahren nach Anspruch 2, **dadurch gekennzeichnet, dass** die Schätzung des Gleichartigkeitskriteriums für das Profil des Pitchs mindestens die folgenden Schritte aufweist:

A1) Auswählen, in der identifizierten Unterklasse des Verzeichnisses der Syntheseeinheiten und ausgehend vom Mittelwert des Pitchs, der im Sinne des Kriteriums des mittleren Pitchs am nächsten liegenden N Einheiten,
A2)zeitliches Abgleichen der N Profile mit demjenigen des zu codierenden Segments,
A3)Berechnen von N Gleichartigkeitsmessungen zwischen den abgeglichenen N Pitch-Profilen und dem Profil des Pitchs des zu codierenden Sprachsegments, um die N Gleichartig-

keitskoeffizienten {rp(1), rp(2), ..., rp(N)} zu erhalten.

5. Verfahren nach Anspruch 2, **dadurch gekennzeichnet, dass** die Gleichartigkeitsschätzung für das energetische Profil mindestens die folgenden Schritte aufweist:

A4)Bestimmen der Entwicklungsprofile der Energie für die ausgewählten N Einheiten gemäß einem Näherungskriterium des mittleren Pitchs,
A5) zeitliches Abgleichen der N Profile mit demjenigen des zu codierenden Segments,
A6)Berechnen von N Gleichartigkeitsmessungen zwischen den abgeglichenen N Energieprofilen und dem Energieprofil des zu codierenden Sprachsegments, um die N Gleichartigkeitskoeffizienten {re(1), re (2), ..., re(N)} zu erhalten.

6. Verfahren nach Anspruch 2, **dadurch gekennzeichnet, dass** die Schätzung der Gleichartigkeitskriterien für die spektrale Hüllkurve mindestens die folgenden Schritte aufweist:

A7)zeitliches Abgleichen der N Profile mit demjenigen des zu codierenden Segments,
A8)Bestimmen der Entwicklungsprofile der spektralen Parameter für die ausgewählten N Einheiten gemäß einem Näherungskriterium des mittleren Pitchs,
A9)Berechnen von N Messungen der Gleichartigkeiten zwischen der spektralen Sequenz des zu codierenden Segments und den dem zu codierenden Sprachsegment entsprechenden extrahierten N spektralen Sequenzen, um die N Gleichartigkeitskoeffizienten {rs(1), rs(2}, ..., rs (N)} zu erhalten.

7. Verfahren nach einem der Ansprüche 4, 5 und 6, **dadurch gekennzeichnet, dass** der zeitliche Abgleich ein durch dynamische Programmierung (DTW) erhaltener zeitlicher Abgleich oder ein Abgleich durch lineare Anpassung der Längen ist.

8. Verfahren nach einem der Ansprüche 4, 5 und 6, **dadurch gekennzeichnet, dass** die Gleichartigkeitsmessung einer genormte Interkorrelationsmessung ist.

9. Verfahren nach Anspruch 6, **dadurch gekennzeichnet, dass** die Gleichartigkeitsmessung eine spektrale Entfernungsmessung ist.

10. Verfahren nach Anspruch 6, **dadurch gekennzeichnet, dass** der Schritt A9) einen Schritt enthält, in dem die Gruppe der Spektren eines gleichen Segments gemittelt wird, und dass die Gleichartigkeitsmessung eine Interkorrelationsmessung ist.

11. Verfahren nach Anspruch 6, **dadurch gekennzeichnet, dass** das spektrale Verzerrungskriterium an mit konstantem Pitch erneut abgetasteten oder mit dem Pitch des zu codierenden Segments erneut abgetasteten harmonischen Strukturen nach Interpolation der anfänglichen harmonischen Strukturen berechnet wird.

12. Verfahren nach einem der Ansprüche 1 bis 11, **dadurch gekennzeichnet, dass** es einen Schritt der Codierung und/oder einen Schritt der Korrektur des Pitchs durch Veränderung des Syntheseprofils aufweist.

13. Verfahren nach Anspruch 12, **dadurch gekennzeichnet, dass** der Schritt der Codierung und/oder der Korrektur des Pitchs eine lineare Umwandlung des Profils des ursprünglichen Pitchs ist.

14. Verwendung des Verfahrens nach einem der Ansprüche 1 bis 12 für die Auswahl und/oder die Codierung von Syntheseeinheiten für einen Sprachcodierer mit sehr niedrigem Datenfluss.

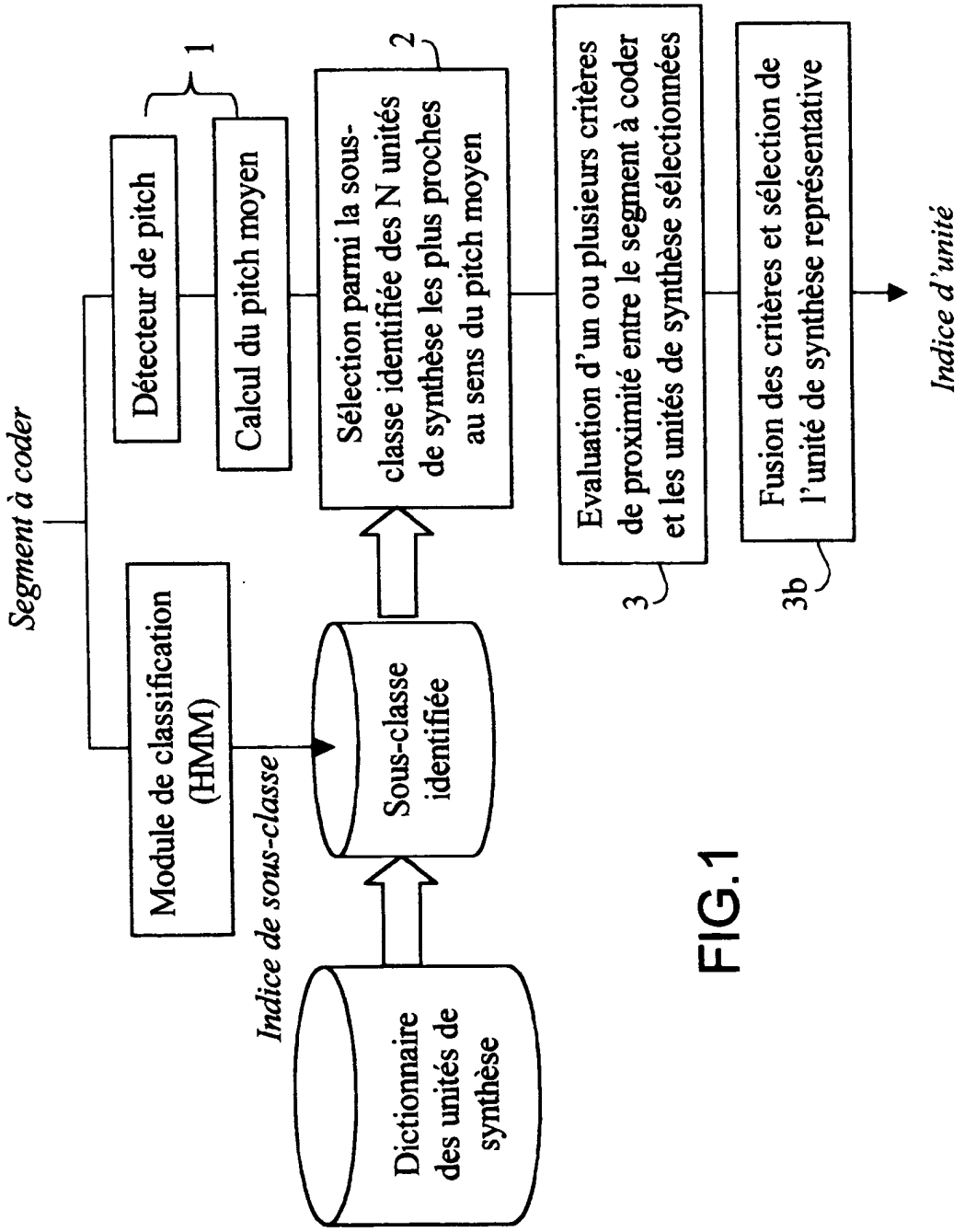
Segment à coder
Détecteur de pitch
Calcul du pitch moyen
1
Module de classification (HMM)
Indice de sous-classe
Sous-classe identifiée
Dictionnaire des unités de synthèse
Sélection parmi la sous-classe identifiée des N unités de synthèse les plus proches au sens du pitch moyen
2
Evaluation d'un ou plusieurs critères de proximité entre le segment à coder et les unités de synthèse sélectionnées
3
Fusion des critères et sélection de l'unité de synthèse représentative
3b
Indice d'unité

FIG.1

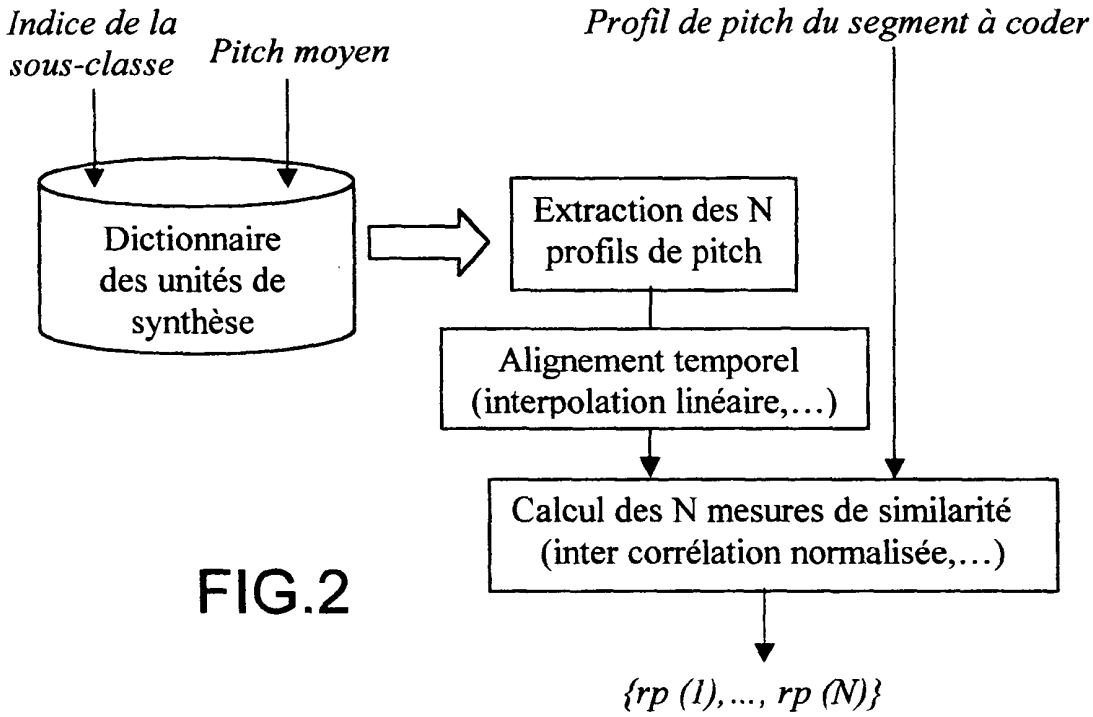
Indice de la sous-classe
Pitch moyen
Profil de pitch du segment à coder
Dictionnaire des unités de synthèse
Extraction des N profils de pitch
Alignement temporel (interpolation linéaire,...)
Calcul des N mesures de similarité (inter corrélation normalisée,...)
{rp (1),..., rp (N)}

FIG.2

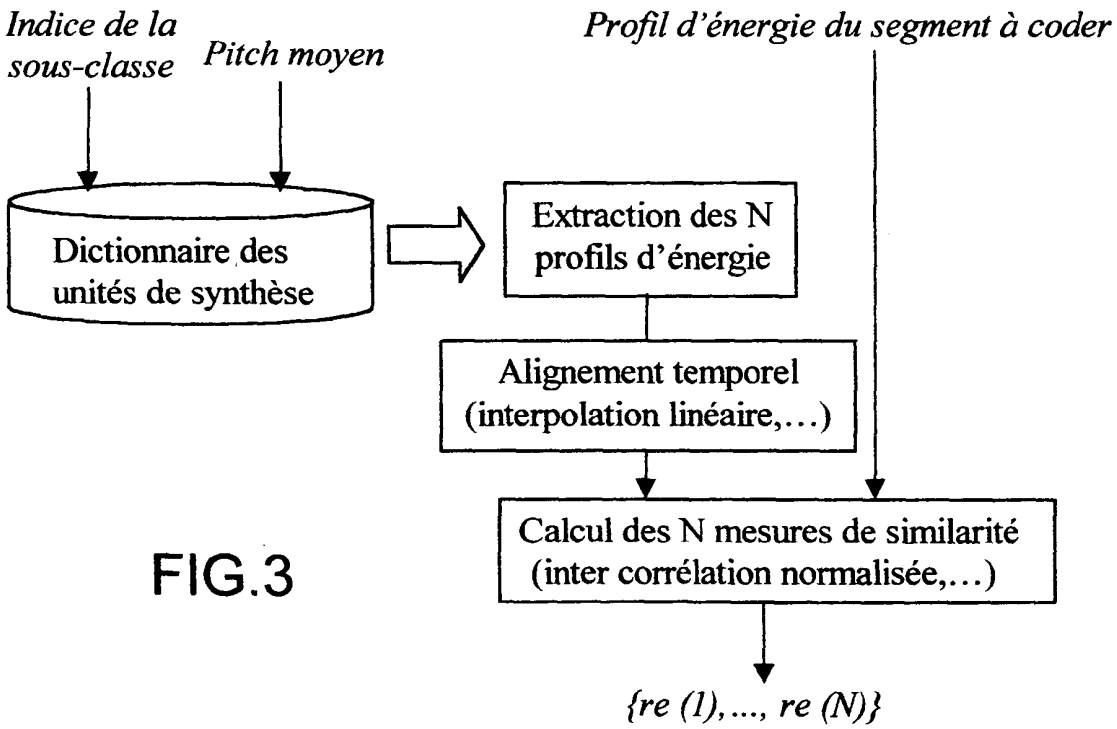
Indice de la sous-classe
Pitch moyen
Profil d'énergie du segment à coder
Dictionnaire des unités de synthèse
Extraction des N profils d'énergie
Alignement temporel (interpolation linéaire,...)
Calcul des N mesures de similarité (inter corrélation normalisée,...)
{re (1),..., re (N)}

FIG.3

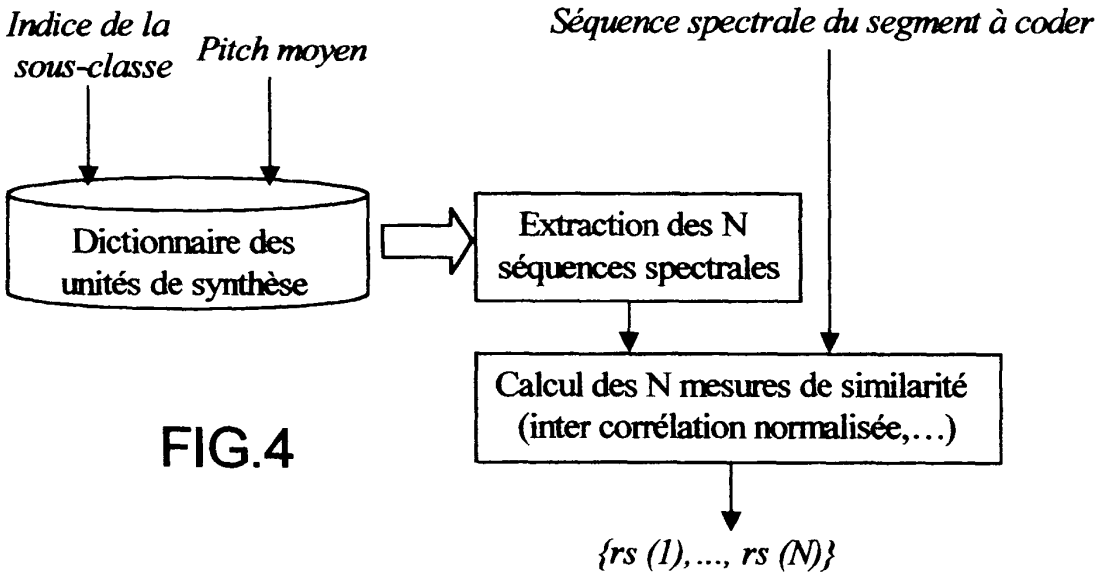
Indice de la sous-classe
Pitch moyen
Séquence spectrale du segment à coder
Dictionnaire des unités de synthèse
Extraction des N séquences spectrales
Calcul des N mesures de similarité (inter corrélation normalisée,...)
{rs (1),..., rs (N)}

FIG.4

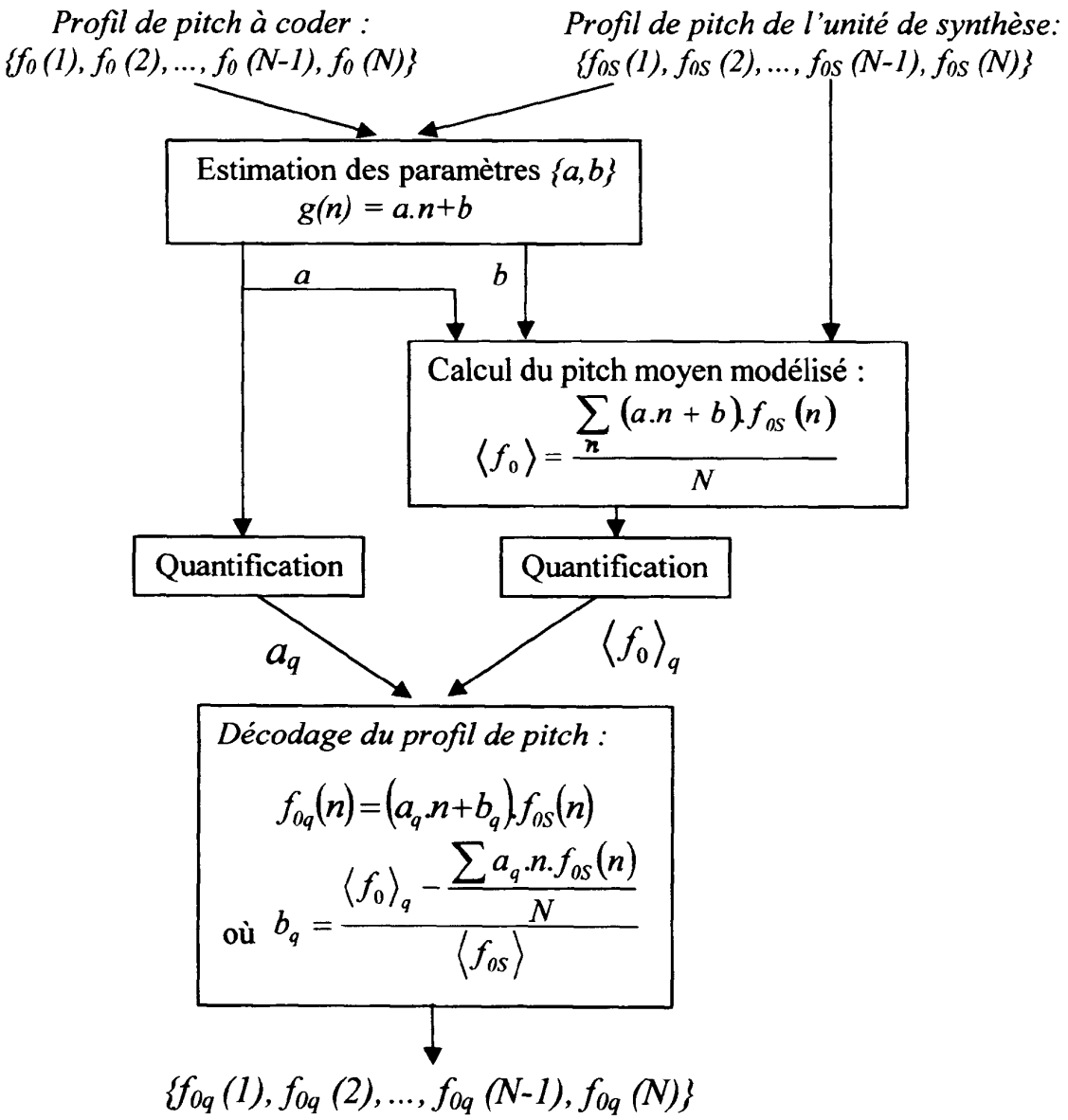
Profil de pitch à coder :
{f0 (1), f0 (2), ..., f0 (N-1), f0 (N)}
Profil de pitch de l'unité de synthèse:
{f0S (1), f0S (2), ..., f0S (N-1), f0S (N)}
Estimation des paramètres {a,b}
g(n) = a.n+b
a
b
Calcul du pitch moyen modélisé :
$\langle f_0 \rangle = \frac{\sum_n (a.n + b).f_{0S}(n)}{N}$
Quantification
Quantification
a_q
$\langle f_0 \rangle_q$
Décodage du profil de pitch :
$f_{0q}(n) = (a_q.n + b_q).f_{0S}(n)$
où $b_q = \frac{\langle f_0 \rangle_q - \frac{\sum a_q.n.f_{0S}(n)}{N}}{\langle f_{0S} \rangle}$
{f0q (1), f0q (2), ..., f0q (N-1), f0q (N)}

FIG.5

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

- **G. Baudoin ; F. El Chami.** Corpus based very low bit rate speech coder. *Proc. Conf. IEEE ICASSP 2003,* 2003 **[0039]**
- **G. Baudoin ; J. Cernocky ; P. Gournay ; G. Chollet.** Codage de la parole à bas et très bas debit. *Annales des télécommunications,* Novembre 2000, vol. 55 (9-10), 421-456 **[0039]**
- **G. Baudoin ; F. Capman ; J. Cernocky ; F. El-chami ; M. Charbit ; G. Chollet ; D. Petrovska-Delacrétaz.** Advances in Very Low Bit Rate Speech Coding using Recognition and Synthesis Techniques. *TSD' 2002,* Septembre 2002, 269-276 **[0039]**
- **K.Lee ; R.Cox.** A segmental coder based on a concatenative TTS. *Speech Communications,* 2002, vol. 38, 89-100 **[0039]**
- **K.Lee ; R.Cox.** A very low bit rate speech coder based on a recognition/synthesis paradigm. *IEEE on ASSP,* Juillet 2001, vol. 9, 482-491 **[0039]**