Office de la Propriété Intellectuelle du Canada
Un organisme d'Industrie Canada

Canadian Intellectual Property Office
An agency of Industry Canada

CA 2359411 C 2010/07/06

(11)(21) 2 359 411

(12) BREVET CANADIEN
CANADIAN PATENT

(13) C

(22) Date de dépôt/Filing Date: 2001/10/17

(41) Mise à la disp. pub./Open to Public Insp.: 2002/04/18

(45) Date de délivrance/Issue Date: 2010/07/06

(30) Priorité/Priority: 2000/10/18 (FR00 13628)

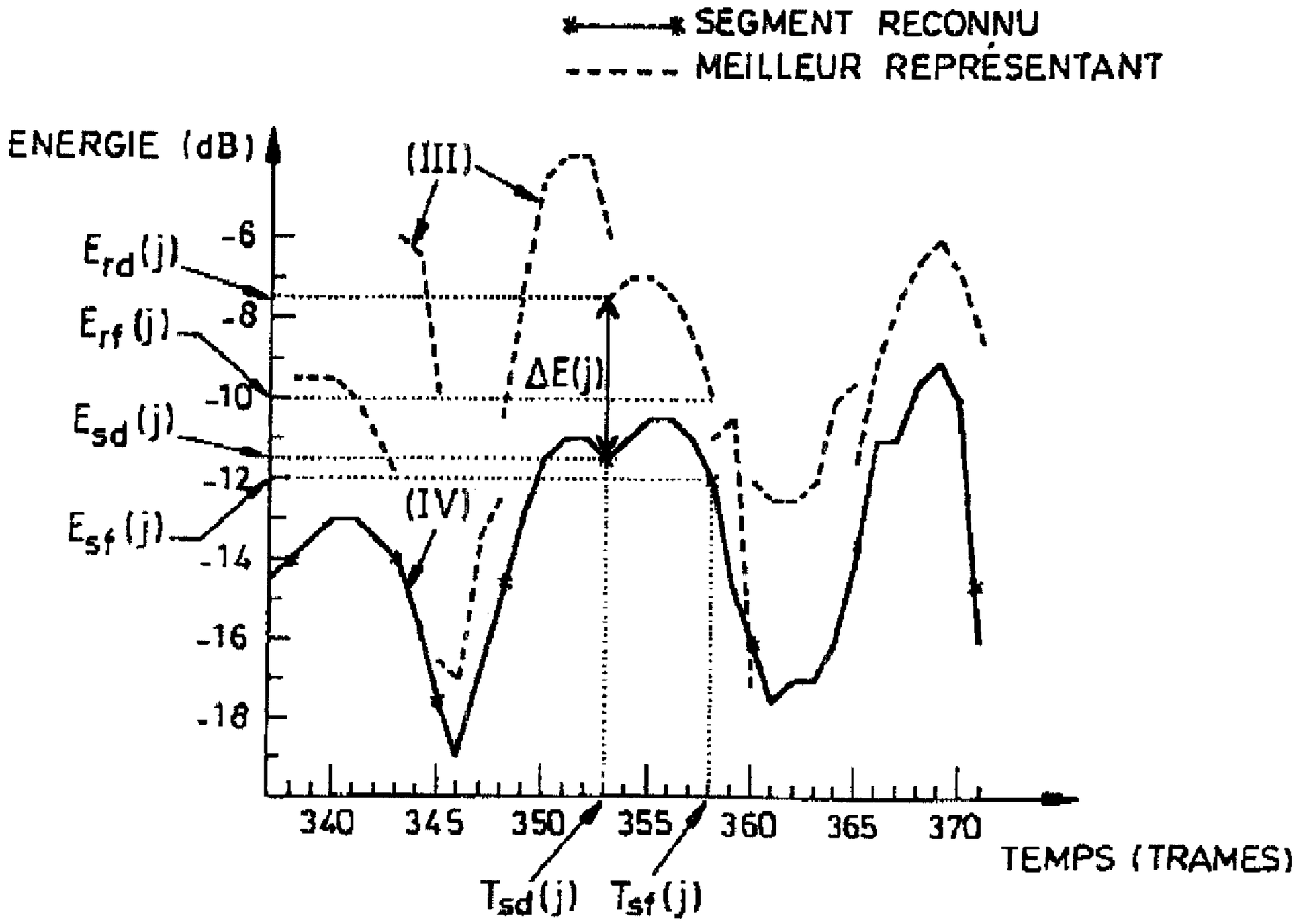
(51) Cl.Int./Int.Cl. *G10L 19/12* (2006.01),
G10L 19/00 (2006.01)

(72) Inventeurs/Inventors:
GOURNAY, PHILIPPE, FR;
NAKACHE, YVES-PAUL, FR

(73) Propriétaire/Owner:
THALES, FR

(74) Agent: GOUDREAU GAGE DUBUC

(54) Titre : PROCEDE DE CODAGE DE LA PROSODIE POUR UN CODEUR DE PAROLE A TRES BAS DEBIT
(54) Title: PROCESS OF CODING OF PROSODY FOR CONVERSATION AT LOW DECIBEL LEVELS



(57) Abrégé/Abstract:
Procédé de codage-décodage de la parole utilisant un codeur à très bas débit comprenant une étape d'apprentissage permettant d'identifier des « représentants » du signal de parole et une étape de codage pour segmenter le signal de parole et déterminer le «

CA 2359411 C 2010/07/06

(11)(21) 2 359 411

(13) C

(57) Abrégé(suite)/Abstract(continued):
meilleur représentant » associé à chaque segment reconnu. Le procédé comporte au moins une étape de codage-décodage d'un des paramètres au moins de la prosodie des segments reconnus, tel que l'énergie et/ou le pitch et/ou le voisement et/ou la longueur des segments, en utilisant une information de prosodie des « meilleurs représentants ». Application pour des débits inférieurs à 400 bits par seconde.

ABRÉGÉ

Procédé de codage-décodage de la parole utilisant un codeur à très bas débit comprenant une étape d'apprentissage permettant d'identifier des « représentants » du signal de parole et une étape de codage pour segmenter le signal de parole et déterminer le « meilleur représentant » associé à chaque segment reconnu. Le procédé comporte au moins une étape de codage-décodage d'un des paramètres au moins de la prosodie des segments reconnus, tel que l'énergie et/ou le pitch et/ou le voisement et/ou la longueur des segments, en utilisant une information de prosodie des « meilleurs représentants ». Application pour des débits inférieurs à 400 bits par seconde.

DOMAINE TECHNIQUE

La présente invention concerne un procédé de codage de la parole à très bas débit et le système associé. Elle s'applique notamment pour des systèmes de codage-décodage de la parole par indexation d'unités de taille variable.

TECHNIQUE ANTÉRIEURE

Le procédé de codage de la parole mis en œuvre à bas débit, par exemple de l'ordre de 2400 bits/s, est généralement celui du vocodeur utilisant un modèle totalement paramétrique du signal de parole. Les paramètres utilisés concernent le voisement qui décrit le caractère périodique ou aléatoire du signal, la fréquence fondamentale des sons voisés encore connue sous le vocable anglo-saxon « PITCH », l'évolution temporelle de l'énergie, ainsi que l'enveloppe spectrale du signal généralement modélisée par un filtre LPC (abréviation anglo-saxonne de Linear Predictive Coding).

Ces différents paramètres sont estimés périodiquement sur le signal de parole, typiquement toutes les 10 à 30 ms. Ils sont élaborés au niveau d'un dispositif d'analyse et sont généralement transmis à distance en direction d'un dispositif de synthèse reproduisant le signal de parole à partir de la valeur quantifiée des paramètres du modèle.

Jusqu'à présent, le plus bas débit normalisé pour un codeur de parole utilisant cette technique est de 800 bits/s. Ce codeur, normalisé en 1994 est décrit par le standard OTAN STANAG 4479 et dans l'article intitulé « NATO STANAG 4479 : A standard for an 800 bps vocoder and channel coding in HF-ECCM system », IEEE Int. Conf. on ASSP, Detroit, pp 480-483, May 1995 ayant pour auteurs Mouy, B., De La Noue, P., et Goudezeune,G. Il repose sur une technique d'analyse trame par trame (22.5 ms) de type LPC 10 et exploite au maximum la redondance temporelle du signal de parole en regroupant les trames 3 par 3 avant encodage des paramètres.

Bien qu'intelligible, la parole reproduite par ces techniques de codage est d'assez mauvaise qualité et n'est plus acceptable à partir du moment où le débit est inférieur à 600 bits/s.

Une manière de réduire le débit est d'utiliser les vocodeurs segmentaux de type phonétiques avec des segments de durée variable qui combinent des principes de reconnaissance et de synthèse de la parole.

La procédure d'encodage utilise essentiellement un système de reconnaissance automatique de la parole en flot continu, qui segmente et « étiquète » le signal de parole selon un nombre d'unités de parole de taille variable. Ces unités phonétiques sont codées par indexation dans un petit dictionnaire. Le décodage repose sur le principe de la synthèse de la parole par concaténation à partir de l'index des unités phonétiques et de la prosodie. Le terme « prosodie » regroupe principalement les paramètres suivants : l'énergie du signal, le pitch, une information de voisement et éventuellement le rythme temporel.

Toutefois, le développement des codeurs phonétiques nécessite des connaissances importantes en phonétique et en liguistique, ainsi qu'une phase de transcription phonétique d'une base de données d'apprentissage qui est coûteuse et qui peut être la source d'erreurs. De plus, les codeurs phonétiques s'adaptent difficilement à une nouvelle langue ou à un nouveau locuteur.

Une autre technique, décrite par exemple dans la thèse de J.Cernocky, intitulée « Speech Processing Using Automatically Derived Segmental Units : Applications to very Low Rate Coding and Speaker Verification » de l'Université Paris XI Orsay, décembre 1998 permet de contourner les problèmes liés à la transcription phonétique de la base de données d'apprentissage en déterminant les unités de parole de façon automatique et indépendamment de la langue.

Le fonctionnement de ce type de codeur se décompose principalement en deux étapes : une étape d'apprentissage et une étape de codage-décodage décrites à la figure 1.

Lors de l'étape d'apprentissage (figure 1), une procédure automatique détermine par exemple après une analyse paramétrique 1 et une étape de segmentation 2, un ensemble de 64 classes d'unités acoustiques désignées « UA ». A chacune de ces classes d'unités acoustiques est associé un modèle statistique 3, de type modèle de Markov (HMM abréviation anglo-saxonne de Hidden Markov Model), ainsi qu'un petit

nombre d'unités représentantes d'une classe, désignées sous le terme « représentants » 4. Dans le système actuel, les représentants sont simplement les 8 unités les plus longues appartenant à une même classe acoustique. Ils peuvent également être déterminés comme étant les N unités plus représentatives de l'unité acoustique. Lors du codage d'un signal de parole après une étape d'analyse paramétrique 5 permettant d'obtenir notamment les paramètres spectraux, les énergies, le pitch, une procédure de reconnaissance (6, 7), à l'aide d'un algorithme de Viterbi, détermine la succession d'unités acoustiques du signal de parole et identifie le « meilleur représentant » à utiliser pour la synthèse de parole. Ce choix se fait par exemple en utilisant un critère de distance spectrale, tel que l'algorithme de DTW (abréviation anglo-saxonne de Dynamic Time Warping).

Le numéro de la classe acoustique, l'indice de cette unité représentante , la longueur du segment, le contenu de DTW et les informations prosodiques issues de l'analyse paramétrique sont transmises au décodeur. La synthèse de la parole se fait par concaténation des meilleurs représentants, éventuellement en utilisant un synthétiseur paramétrique de type LPC.

Pour concaténer les représentants lors du décodage de la parole, on fait appel, par exemple, à un procédé d'analyse/synthèse paramétrique de la parole. Ce procédé paramétrique permet notamment des modifications de prosodie telles que l'évolution temporelle, la fréquence fondamentale ou pitch, par rapport à une simple concaténation de formes d'onde.

Le modèle paramétrique de parole utilisé par le procédé d'analyse/synthèse peut être à excitation binaire voisé/ non voisé de type LPC 10 tel que décrit dans le document intitulé « The government standard linear predictive coding algorithm : LPC-10 » de T.Tremain publié dans la revue Speech Technology, vol.1, n°2, pp 40-49.

Cette technique permet de coder l'enveloppe spectrale du signal en 185 bits/s environ pour un système monolocuteur, pour une moyenne d'environ 21 segments par seconde.

Dans la suite de la description les termes ci-après ont les significations suivantes :

- le terme « représentant » correspond à l'un des segments de la base d'apprentissage qui a été jugé représentatif d'une des classes d'unités acoustique,

- l'expression « segment reconnu » correspond à un segment de la parole qui a été identifié comme appartenant à l'une des classes acoustiques, par le codeur,
- l'expression « meilleur représentant » désigne le représentant déterminé au niveau du codage qui représente le mieux le segment reconnu.

RÉSUMÉ DE L'INVENTION

L'objet de la présente invention concerne un procédé de codage, décodage de la prosodie pour un codeur de parole à très bas débit utilisant notamment les meilleurs représentants.

Il concerne aussi la compression de données.

L'invention concerne un procédé de codage-décodage de la parole utilisant un codeur à très bas débit comprenant une étape d'apprentissage permettant d'identifier des « représentants » du signal de parole et une étape de codage pour segmenter le signal de parole et déterminer le « meilleur représentant » associé à chaque segment reconnu. Il est caractérisé en ce qu'il comporte au moins une étape de codage-décodage d'un des paramètres au moins de la prosodie des segments reconnus, tel que l'énergie et/ou le pitch et/ou le voisement et/ou la longueur des segments, en utilisant une information de prosodie des « meilleurs représentants ».

L'information de prosodie des représentants utilisée est par exemple le contour d'énergie ou le voisement ou la longueur des segments ou le pitch. .

L'étape de codage de la longueur des segments reconnus consiste par exemple à coder la différence de longueur entre la longueur d'un segment reconnu et la longueur du « meilleur représentant » multiplié par un facteur donné.

Selon un mode de réalisation, il comporte une étape de codage de l'alignement temporel des meilleurs représentants en utilisant le chemin de DTW et en recherchant le plus proche voisin dans une table de formes.

L'étape de codage de l'énergie peut comporter une étape de détermination pour chaque début de « segment reconnu » de la différence $\Delta E(j)$ entre la valeur d'énergie $E_{rd}(j)$ du « meilleur représentant » et la valeur

d'énergie $E_{sd}(j)$ du début du « segment reconnu » et l'étape de décodage comporter pour chaque segment reconnu, une première étape consistant à translater le contour d'énergie du meilleur représentant d'une quantité $\Delta E(j)$ pour faire coïncider la première énergie $E_{rd}(j)$ du « meilleur représentant » avec la première énergie $E_{sd}(j+1)$ du segment reconnu d'indice j+1.

L'étape de codage de voisement comporte par exemple une étape de détermination des différences existantes ΔT_k pour chaque extrémité d'une zone de voisement d'indice k entre la courbe du voisement des segments reconnus et celle des meilleurs représentants et l'étape de décodage comporte par exemple pour chaque extrémité d'une zone de voisement d'indice k une étape de correction de la position temporelle de cette extrémité d'une valeur ΔT_k correspondante et/ou une étape de suppression ou d'insertion d'une transition.

Le procédé concerne aussi un système de codage-décodage de la parole comportant au moins une mémoire pour stocker un dictionnaire comprenant un ensemble de représentants du signal de parole, un microprocesseur adapté pour déterminer les segments reconnus, pour reconstruire la parole à partir des « meilleurs représentants » et pour mettre en œuvre les étapes du procédé selon l'une des caractéristiques précitées.

Le dictionnaire des représentants est par exemple commun au codeur et au décodeur du système codage-décodage.

Le procédé et le système selon l'invention peuvent être utilisés pour le codage-décodage de la parole pour des débits inférieurs à 800 bits/s et de préférence inférieurs à 400 bits/s.

Le procédé et le système de codage-décodage selon l'invention offrent notamment l'avantage de coder à très bas débit la prosodie et de fournir ainsi un codeur complet dans ce domaine d'application.

BRÈVE DESCRIPTION DES DESSINS

D'autres caractéristiques et avantages apparaîtront à la lecture de la description détaillée d'un mode de réalisation pris à titre d'exemple non limitatif et illustré par les dessins annexés où :

- la figure 1 représente un schéma d'apprentissage, de codage et de décodage de la parole selon l'art antérieur,

- les figures 2 et 3 décrivent des exemples de codage de la longueur des segments reconnus,
- la figure 4 schématise un modèle d'alignement temporel des « meilleurs représentants »,
- les figures 5 et 6 montrent des courbes des énergies du signal à coder et des représentants alignés, ainsi que les contours des énergies initial et décodé obtenus en mettant en œuvre le procédé selon l'invention,
- la figure 7 schématise le codage du voisement du signal de parole, et
- la figure 8 est un exemple de codage du pitch.

DESCRIPTION DÉTAILLÉE

Le principe de codage selon l'invention repose sur l'utilisation des « meilleurs représentants », notamment leur information de prosodie, pour coder et/ou décoder au moins un des paramètres de prosodie d'un signal de parole, par exemple le pitch, l'énergie du signal, le voisement, la longueur des segments reconnus.

Pour compresser la prosodie à très bas débit, le principe mis en œuvre utilise la segmentation du codeur ainsi que les informations prosodiques des « meilleurs représentants ».

La description qui suit donnée à titre illustratif et nullement limitatif décrit un procédé de codage de la prosodie dans un dispositif de codage-décodage de la parole à faible débit qui comporte un dictionnaire obtenu de façon automatique, par exemple, lors de l'apprentissage tel que décrit à la figure 1.

Le dictionnaire comprend les informations suivantes :

- plusieurs classes d'unités acoustiques UA, chaque classe étant déterminée à partir d'un modèle statistique,
- pour chaque classe d'unités acoustiques, un ensemble de représentants.

Ce dictionnaire est connu du codeur et du décodeur. Il correspond par exemple à une ou plusieurs langues et à un ou plusieurs locuteurs.

Le système de codage-décodage comporte par exemple une mémoire pour stocker le dictionnaire, un microprocesseur adapté pour déterminer les segments reconnus, pour la mise en œuvre des différentes

étapes du procédé selon l'invention et pour reconstruire la parole à partir des meilleurs représentants.

Le procédé selon l'invention met œuvre au moins une des étapes suivantes : le codage de la longueur des segments, le codage de l'alignement temporel des « meilleurs représentants », le codage et/ou le décodage de l'énergie, le codage et/ou le décodage de l'information de voisement et/ou le codage et/ou le décodage du pitch et/ou le décodage de la longueur des segments et de l'alignement temporel.

Codage de la longueur des segments

Le système de codage détermine en moyenne un nombre Ns de segments par seconde, par exemple 21 segments. La taille de ces segments varie en fonction de la classe d'unités acoustiques UA. Il apparaît que pour la majorité des UA, le nombre de segments décroît selon une relation $1/x^{2.6}$, où x est la longueur du segment.

Une variante de réalisation du procédé selon l'invention consiste à coder la différence de longueur variable entre le « segment reconnu » et la longueur du « meilleur représentant » selon un schéma décrit à la figure 2.

Sur ce schéma dans la colonne de gauche figure la longueur du mot de code à utiliser et dans la colonne de droite la différence de longueur entre la longueur du segment reconnu par le codeur pour le signal de parole et celle du meilleur représentant.

Selon un autre mode de réalisation donnée à la figure 3, le codage de la longueur absolue d'un segment reconnu est effectué à l'aide d'un code à longueur variable semblable à celui de Huffman connu de l'Homme du métier, ce qui permet d'obtenir un débit de l'ordre de 55 bits/s.

Le fait d'utiliser les longs mots de code pour coder les longueurs de grands segments reconnus, permet notamment de conserver la valeur de débit dans une plage de variation limitée. En effet, ces longs segments réduisent le nombre de segment reconnu par seconde et le nombre de longueurs à coder.

En résumé, on code par exemple avec un code à longueur variable la différence entre la longueur du segment reconnu et la longueur du meilleur représentant multiplié par un certain facteur, ce facteur pouvant être compris entre 0 (codage absolu) et 1 (codage de la différence).

Codage de l'alignement temporel des meilleurs représentants

L'alignement temporel est par exemple réalisé en suivant le chemin de la DTW (abréviation anglo-saxonne de Dynamic Time Warping) qui a été déterminé lors de la recherche du « meilleur représentant » pour coder le « segment reconnu ».

La figure 4 représente le chemin (C) de la DTW correspondant au contour temporel qui minimise la distorsion entre le paramètre à coder (axe des abscisses), par exemple le vecteur des coefficients « cepstraux », et le « meilleur représentant » (axe des ordonnées). Cette approche est décrite dans le livre ayant pour titre « Traitement de la parole », pour auteur René Boite et Murat Kunt publié aux Presses Polytechnique Romandes éditions 1987.

Le codage de l'alignement des « meilleurs représentants » est effectué par recherche du plus proche voisin dans une table contenant des formes type. Le choix de ces formes type se fait par exemple par une approche statistique, telle que l'apprentissage sur une base de données de parole ou par une approche algébrique par exemple la description par des équations mathématiques paramétrables, ces différentes méthodes étant connues de l'Homme du métier.

Selon une autre approche, valable dans le cas où les segments de petite taille sont en proportion importante, le procédé effectue un alignement des segments suivant la diagonale plutôt que le chemin exact de la DTW. Le débit est alors nul.

Codage-décodage de l'énergie

Lorsque l'on classe et analyse les segments de la base de données de parole appartenant à chacune des classes d'unités acoustiques, on constate qu'il se dégage une certaine cohérence dans la forme des contours des énergies. De plus, il existe des ressemblances entre les contours d'énergie des meilleurs représentants alignés par DTW et les contours de l'énergie du signal à coder.

Le codage de l'énergie est décrit ci-après en relation aux figures 5 et 6, où l'axe des ordonnées correspond à l'énergie du signal de la parole à coder exprimée en dB et l'axe des abscisses au temps exprimé en trames.

La figure 5 représente la courbe (III) regroupant des contours d'énergie des meilleurs représentants alignés et la courbe (IV) des contours d'énergie des segments reconnus séparés par des * sur la figure. Un segment reconnu d'indice j est délimité par deux points de coordonnées respectives $[E_{sd}(j) ; T_{sd}(j)]$ et $[E_{sf}(j) ; T_{sf}(j)]$ où $E_{sd}(j)$ est l'énergie de début de segment et $E_{sf}(j)$ l'énergie de fin de segment, pour les instants T_{df} et T_{sf} correspondant. Les références $E_{rd}(j)$ et $E_{rf}(j)$ sont utilisées pour les valeurs d'énergies du début et de la fin d'un « meilleur représentant » et la référence $\Delta E(j)$ correspond à la translation déterminée pour un segment reconnu d'indice j.

Codage de l'énergie

Le procédé comporte une première étape de détermination de la translation à réaliser.

Pour cela on détermine pour chaque début de « segment reconnu », la différence $\Delta E(j)$ existant entre la valeur d'énergie $E_{rd}(j)$ du meilleur représentant (courbe III) et la valeur d'énergie E_{sd} du début du segment reconnu (courbe IV). On obtient un ensemble de valeurs $\Delta E(j)$ que l'on quantifie par exemple uniformément de manière à connaître la translation à appliquer lors du décodage. La quantification est réalisée par exemple en utilisant des méthodes connues de l'Homme du métier.

Décodage de l'énergie du signal de parole

Le procédé consiste notamment à utiliser les contours d'énergie des meilleurs représentants (courbe III) pour reconstruire les contours d'énergie du signal à coder (courbe IV).

Pour chaque segment reconnu, une première étape consiste à translater le contour d'énergie du meilleur représentant pour la faire coïncider avec la première énergie $E_{rd}(j)$ en lui appliquant la translation $\Delta E(j)$, définie à l'étape de codage par exemple, pour déterminer la valeur $E_{sd}(j)$. Après cette première étape de translation, le procédé comporte une étape de modification de la pente du contour d'énergie du meilleur représentant afin de relier la dernière valeur d'énergie $E_{rd}(j)$ du « meilleur représentant » à la première énergie $E_{sd}(j+1)$ du segment suivant d'indice j+1.

La figure 6 représente les courbes (VI) et (VII) correspondant respectivement au contour d'énergie original du signal de parole à coder et

du contour d'énergie décodé après mise en œuvre des étapes décrites précédemment.

Par exemple, le codage des énergies de début de chaque segment sur 4 bits permet d'obtenir pour le codage segmental de l'énergie un débit de l'ordre de 80 bits/s.

Codage de l'information de voisement

La figure 7 représente l'évolution temporelle d'une information de voisement binaire de quatre segments successifs 35, 36, 37 pour le signal à coder courbe (VII) et pour les meilleurs représentants (courbe VIII) après alignement temporel par DTW.

Codage de l'information de voisement

Lors du codage, le procédé exécute une étape de codage de l'information de voisement, par exemple en parcourant l'évolution temporelle de l'information de voisement des segments reconnus et celle des meilleurs représentants alignés (courbe VIII) et en codant les différences existantes ΔT_k entre ces deux courbes. Ces différences ΔT_k peuvent être : une avance a de la trame, un retard b de trame, l'absence et/ou la présence d'une transition référence c (k correspond à l'indice d'une extrémité d'une zone de voisement).

Pour cela, il est possible d'utiliser un code de longueur variable dont un exemple est donné dans la table I ci-dessous, pour coder la correction à apporter à chacune des transitions de voisement pour chacun des segments reconnus. Tous les segments ne comportant pas de transition de voisement, il est possible de réduire le débit associé au voisement en ne codant que les transitions de voisement existantes dans le voisement à coder et dans les meilleurs représentants.

Selon cette méthode, l'information de voisement est codée sur environ 22 bits par seconde.

Table 1 : *Exemple de table de codage pour les transitions de voisement :*

Code	*Interprétation*
000	Transition à supprimer
001	Décalage 1 trame à Droite
010	Décalage 1 trame à Gauche
011	Décalage 2 trames à Droite
100	Décalage 2 trames à Gauche
101	Insérer une transition (un code précisant l'emplacement de la transition suit celui-ci)
110	Pas de décalage
111	Déplacement supérieur à 3 trames (un autre code suit celui-ci)

Pour une information de voisement mixte telle que :

- le taux de voisement en sous-bande, l'analyse de cette information fait appel à une méthode décrite par exemple dans le document suivant : "Multiband Excitation Vocoders", ayant pour auteurs D.W. Griffin and J.S. Lim, IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 36, no. 8, pp. 1223-1235, 1988 ;
- la fréquence de transition entre une bande basse voisée et une bande haute non-voisée, le codage utilise une méthode telle que décrite dans le document ayant pour auteurs C. Laflamme, R. Salami, R. Matmti, and J-P. Adoul, intitulé "Harmonic Stochastic Excitation (HSX) speech coding below 4 kbits/s", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, May 1996, pp. 204-207.

Dans ces deux cas, le codage de l'information de voisement comporte également le codage de la variation de la proportion de voisement.

<u>Décodage de l'information de voisement</u>

Le décodeur dispose de l'information de voisement des « meilleurs représentants alignés » obtenu au niveau du codeur.

La correction s'effectue par exemple de la manière suivante :

A chaque détection de l'extrémité d'une zone de voisement sur les meilleurs représentants choisis pour la synthèse, le procédé apporte une information complémentaire au décodeur qui est la correction à effectuer à cette extrémité. La correction peut être une avance a ou un retard b à apporter à cette extrémité. Ce décalage temporel est par exemple exprimé

en nombre de trames afin d'obtenir la position exacte de l'extrémité de voisement du signal de parole original. La correction peut aussi prendre la forme d'une suppression ou d'une insertion d'une transition.

Codage du pitch

L'expérience montre que, sur des enregistrements de parole, le nombre de zones voisées obtenues par seconde est en moyenne de l'ordre de 3 ou 4. Pour rendre compte fidèlement des variations du pitch, une manière de procéder consiste à transmettre plusieurs valeurs de pitch par zone voisée. Afin de limiter le débit, au lieu de transmettre toute la succession des valeurs de pitch sur une zone voisée, le contour du pitch est approximé par une succession de segments linéaires.

Codage du pitch

Pour chaque zone voisée du signal de parole, le procédé comporte une étape de recherche des valeurs du pitch à transmettre. Les valeurs de pitch au début et à la fin de la zone voisée sont systématiquement transmises. Les autres valeurs à transmettre sont déterminées de la manière suivante :

- le procédé considère uniquement les valeurs du pitch au début des segments reconnus. Partant de la droite Di joignant les valeurs du pitch aux deux extrémités de la zone voisée, le procédé recherche le début de segment dont la valeur de pitch est la plus éloignée de cette droite, ce qui correspond à une distance d_{max}. Il compare cette valeur d_{max} à une valeur seuil d_{seuil}. Si la distance d_{max} est supérieure à d_{seuil}, le procédé décompose la droite initiale Di en deux droites D_{i1} et D_{i2}, en prenant le début du segment trouvé comme nouvelle valeur de pitch à transmettre. Cette opération est réitérée sur ces deux nouvelles zones voisée délimitées par les droites D_{i1} et D_{i2} jusqu'à ce que la distance d_{max} trouvée soit inférieure à la distance d_{seuil}.

Pour coder les valeurs du pitch ainsi déterminées, le procédé utilise par exemple un quantificateur scalaire prédictif sur par exemple 5 bits appliqué au logarithme du pitch.

La prédiction est par exemple la première valeur de pitch du meilleur représentant correspondant à la position du pitch à décoder, multipliée par un facteur de prédiction compris par exemple entre 0 et 1.

Selon une autre façon de procéder, la prédiction peut être la valeur minimale de l'enregistrement de parole à coder. Dans ce cas, cette valeur peut être transmise au décodeur par quantification scalaire sur par exemple 8 bits.

Les valeurs des pitchs à transmettre ayant été déterminées et codées, le procédé comporte une étape où l'espacement temporel est précisé, par exemple en nombre de trames, entre chacune de ces valeurs de pitch. Un code à longueur variable permet par exemple de coder ces espacements sur 2 bits en moyenne.

Cette façon de procéder permet d'obtenir un débit d'environ 65/bits par seconde pour une distance maximale sur la période pitch de 7 échantillons.

Décodage du pitch

L'étape de décodage comporte tout d'abord une étape de décodage de l'espacement temporel entre les différentes valeurs de pitch transmises afin de récupérer les instants de mise à jour du pitch, ainsi que la valeur du pitch pour chacun de ces instants. La valeur du pitch pour chacune des trames de la zone voisée est reconstituée par exemple par interpolation linéaire entre les valeurs transmises.

REVENDICATIONS

1. Une méthode de codage de la parole comprenant:

une étape d'apprentissage comportant

des représentants d'apprentissage d'un premier signal de parole, chaque représentant étant stocké dans une base de données parmi un ensemble de représentants d'une classe d'unités acoustiques, chaque classe d'unités acoustiques étant basée sur un modèle statistique et non sur des phonèmes ou des mots prédéterminés;

une étape de codage comportant

une étape de segmentation d'un deuxième signal de parole,

une étape de détermination des segments reconnus du deuxième signal de parole, chaque segment reconnu comportant une portion du deuxième signal de parole correspondant à au moins un des représentants stockés dans la base de données,

une étape de détermination des meilleurs représentants associés à au moins une information de prosodie des segments reconnus, chaque meilleur représentant étant choisi parmi les représentants de la même classe d'unités acoustiques comme étant le représentant qui s'approche le mieux de ladite au moins une information de prosodie du segment reconnu correspondant, et

une étape de codage du deuxième signal de parole à un débit inférieur à 800 bits/s en codant au moins un premier meilleur représentant de ladite au moins une information de prosodie d'un premier segment reconnu correspondant et en codant une différence entre ladite au moins une information de prosodie du premier meilleur représentant et ladite au moins une information de prosodie du premier segment reconnu;

une étape de codage d'un alignement temporel des meilleurs représentants en utilisant un chemin de la Dynamic Time Warping (DTW); et

une étape de recherche du plus proche voisin dans une table de formes.

2. Méthode selon la revendication 1, dans laquelle ladite au moins une information de prosodie est une énergie, un voisement, une longueur, ou un pitch du premier segment reconnu du signal de parole et du premier meilleur représentant.

3. Méthode selon la revendication 2, dans laquelle le codage de la différence entre ladite au moins une information de prosodie du premier meilleur représentant et du premier segment reconnu comprend une étape de codage de longueur comportant:

une étape de codage d'une différence de longueur entre une longueur du premier segment reconnu et une longueur du premier meilleur représentant; et

une étape de multiplication de la différence de longueur par un facteur donné.

4. Méthode selon la revendication 2, dans laquelle le codage de la différence entre ladite au moins une information de prosodie du premier meilleur représentant et du premier segment reconnu comprend une étape de codage d'énergie comportant:

une étape de détermination d'une différence $\Delta E(j)$ entre une valeur d'énergie $E_{rd}(j)$ d'un début du meilleur représentant et une valeur d'énergie $E_{sd}(j)$ d'un début du premier segment reconnu.

5. Méthode selon la revendication 4, comprenant une étape de décodage d'énergie comportant:

une étape de translation d'un contour d'énergie du premier meilleur représentant de la quantité $\Delta E(j)$ pour faire coïncider la valeur d'énergie $E_{rd}(j)$ du début du premier meilleur représentant avec la valeur d'énergie $E_{sd}(j)$ du début du premier segment reconnu, et

une étape de modification de la pente du contour d'énergie du premier meilleur représentant pour faire coïncider une dernière valeur d'énergie

$E_{rd}(j)$ du premier meilleur représentant avec une valeur d'énergie $E_{sd}(j+1)$ d'un début d'un segment reconnu d'indice j+1.

6. Méthode selon la revendication 2, dans laquelle le codage de la différence entre ladite au moins une information de prosodie du premier meilleur représentant et du premier segment reconnu comprend une étape de codage d'une information de voisement comportant:

une étape de détermination d'une différence ΔT_k, pour une extrémité d'une zone de voisement d'indice k, entre les courbes de voisement du premier segment reconnu et du premier meilleur représentant.

7. Méthode selon la revendication 6, comprenant une étape de décodage de l'information de voisement comportant:

une étape de correction, pour l'extrémité de la zone de voisement d'indice k, d'une position temporelle de cette extrémité d'une valeur ΔT_k; ou

une étape de suppression ou d'insertion d'une transition.

8. Méthode selon la revendication 1, dans laquelle l'étape de codage du deuxième signal de parole est à un débit inférieur à 400 bits/s.

9. Méthode selon la revendication 1, dans laquelle le codage de la différence entre ladite au moins une information de prosodie du premier meilleur représentant et du premier segment reconnu comprend une étape de codage de pitch comportant:

(a) une étape d'estimation d'un contour de pitch d'une zone voisée en formant une droite D_i joignant une valeur de pitch au début d'un premier segment reconnu à une valeur de pitch au début d'un segment reconnu suivant;

(b) une étape de détermination d'une distance d_{max} la plus grande entre la droite et le contour de pitch;

(c) une étape de comparaison de la plus grande distance d_{max} avec une valeur seuil de distance d_{seuil} prédéterminée; et

(d) si la plus grande distance d_{max} est supérieure à la valeur seuil d_{seuil} prédéterminée, une étape de division de la zone voisée en une première

zone voisée qui s'étend du début du premier segment reconnu à la valeur de pitch qui définit la plus grande distance d_{max} et une deuxième zone voisée qui s'étend de la valeur de pitch qui définit la plus grande distance d_{max} au début du segment reconnu suivant.

10. Un système de codage de signal de parole comprenant:

un codeur comportant

une unité configurée pour apprendre des représentants d'un premier signal de parole, chaque représentant étant stocké dans une base de données parmi un ensemble de représentants d'une classe d'unités acoustiques, chaque classe d'unités acoustiques étant basée sur un modèle statistique et non sur des phonèmes ou des mots prédéterminés,

une unité adaptée à segmenter un deuxième signal de parole,

une unité configurée pour déterminer des segments reconnus du deuxième signal de parole, chaque segment reconnu comportant une portion du deuxième signal de parole correspondant à au moins un des représentants stockés dans la base de données,

une unité adaptée à déterminer les meilleurs représentants associés à au moins une information de prosodie des segments reconnus, chaque meilleur représentant étant choisi parmi les représentants de la même classe d'unités acoustiques comme étant le représentant qui s'approche le mieux de ladite au moins une information de prosodie du segment reconnu correspondant, et

une unité adaptée à coder le deuxième signal de parole à un débit inférieur à 800 bits/s en codant un premier meilleur représentant de ladite au moins une information de prosodie d'un premier segment reconnu correspondant et en codant une différence entre ladite au moins une information de prosodie du premier meilleur représentant et ladite au moins une information de prosodie du premier segment reconnu; et

au moins une mémoire adaptée à stocker la base de données des représentants.

11. Système selon la revendication 10, comprenant:

un décodeur,

la mémoire adaptée à stocker la base de données des représentants étant commune au codeur et au décodeur du système de codage.

12. Système selon la revendication 10, dans lequel le codeur est adapté à coder le deuxième signal de parole à un débit inférieur à 400 bits/s.

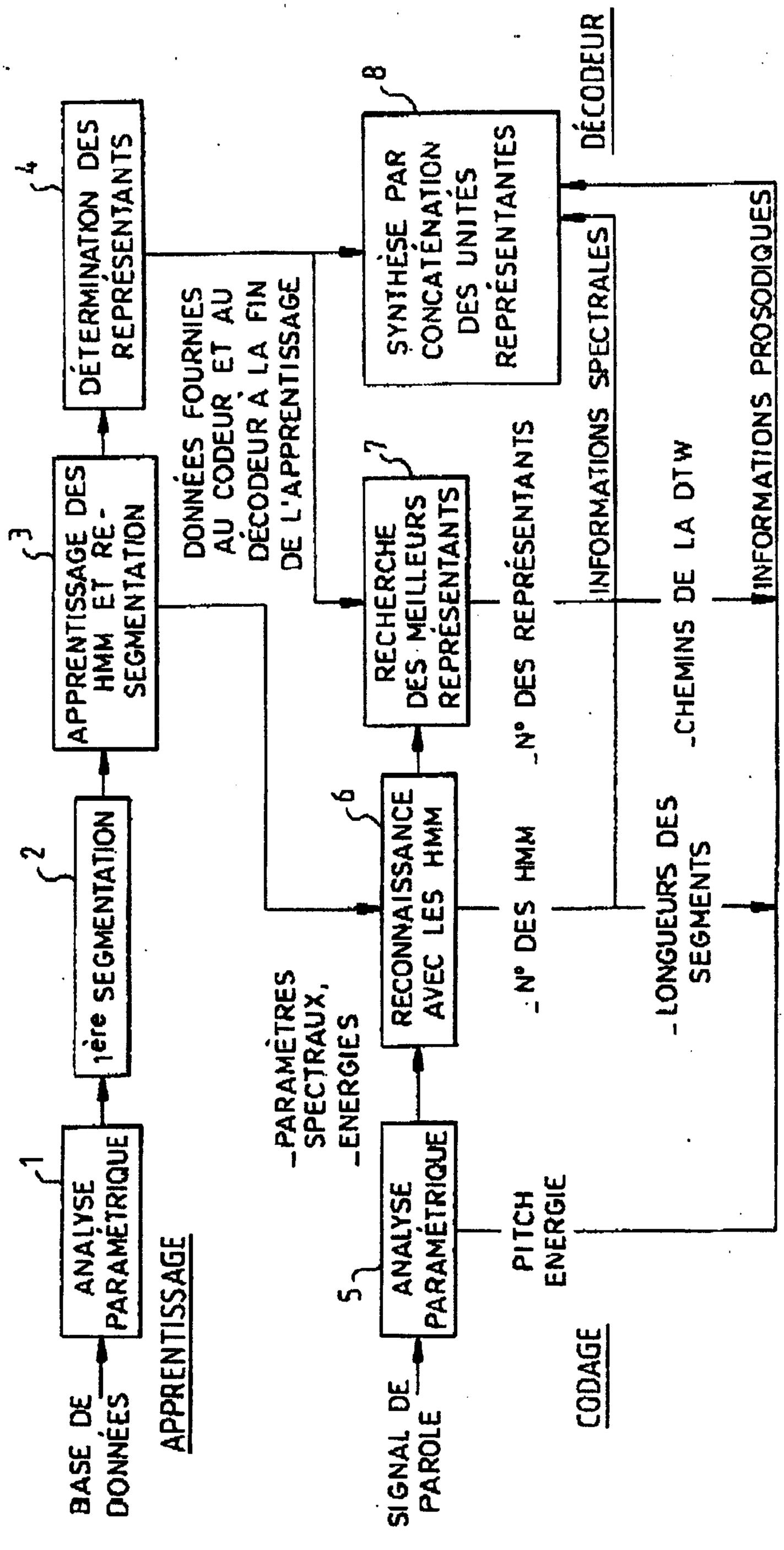


FIG.1

ART ANTÉRIEUR

FIG.2

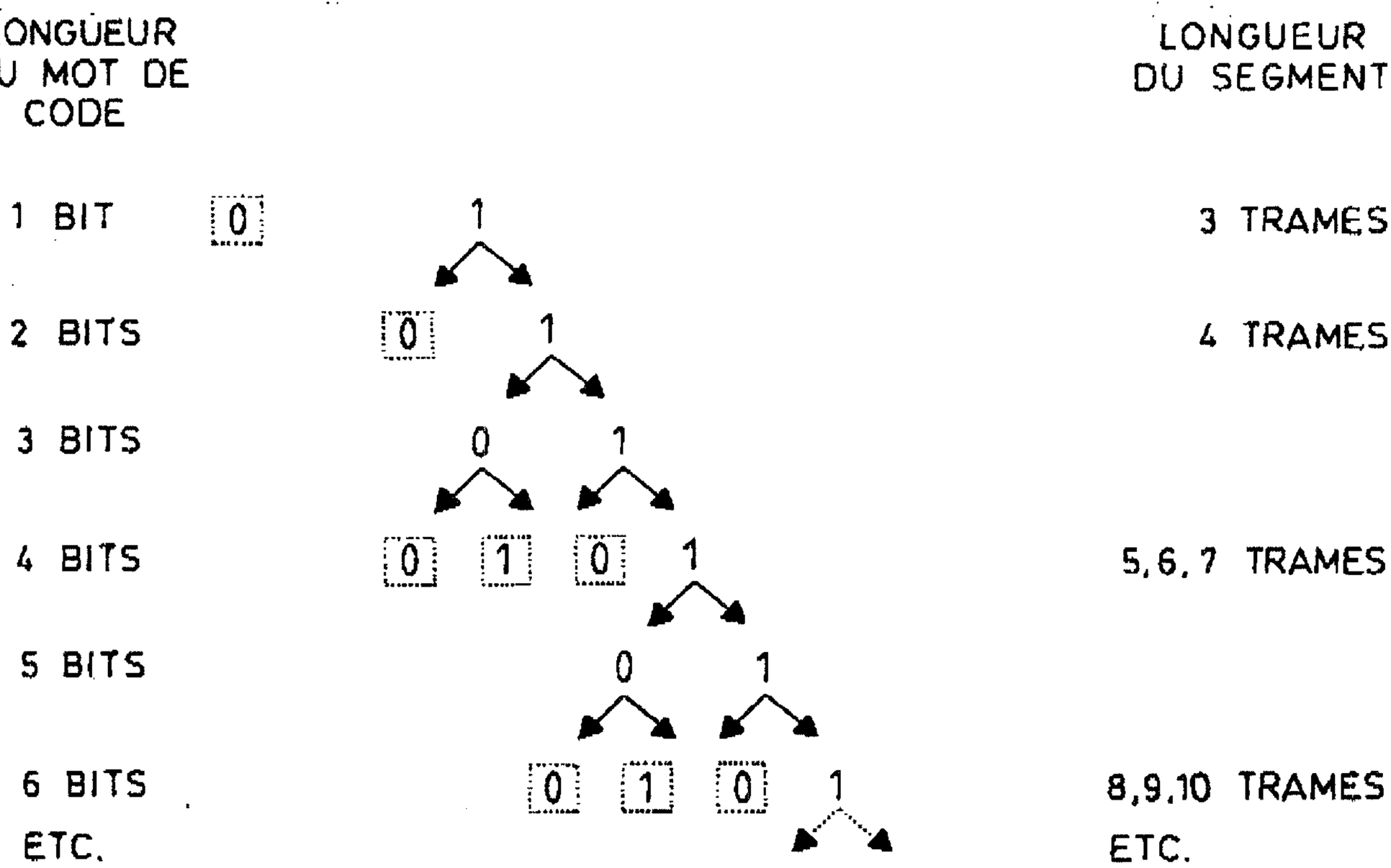
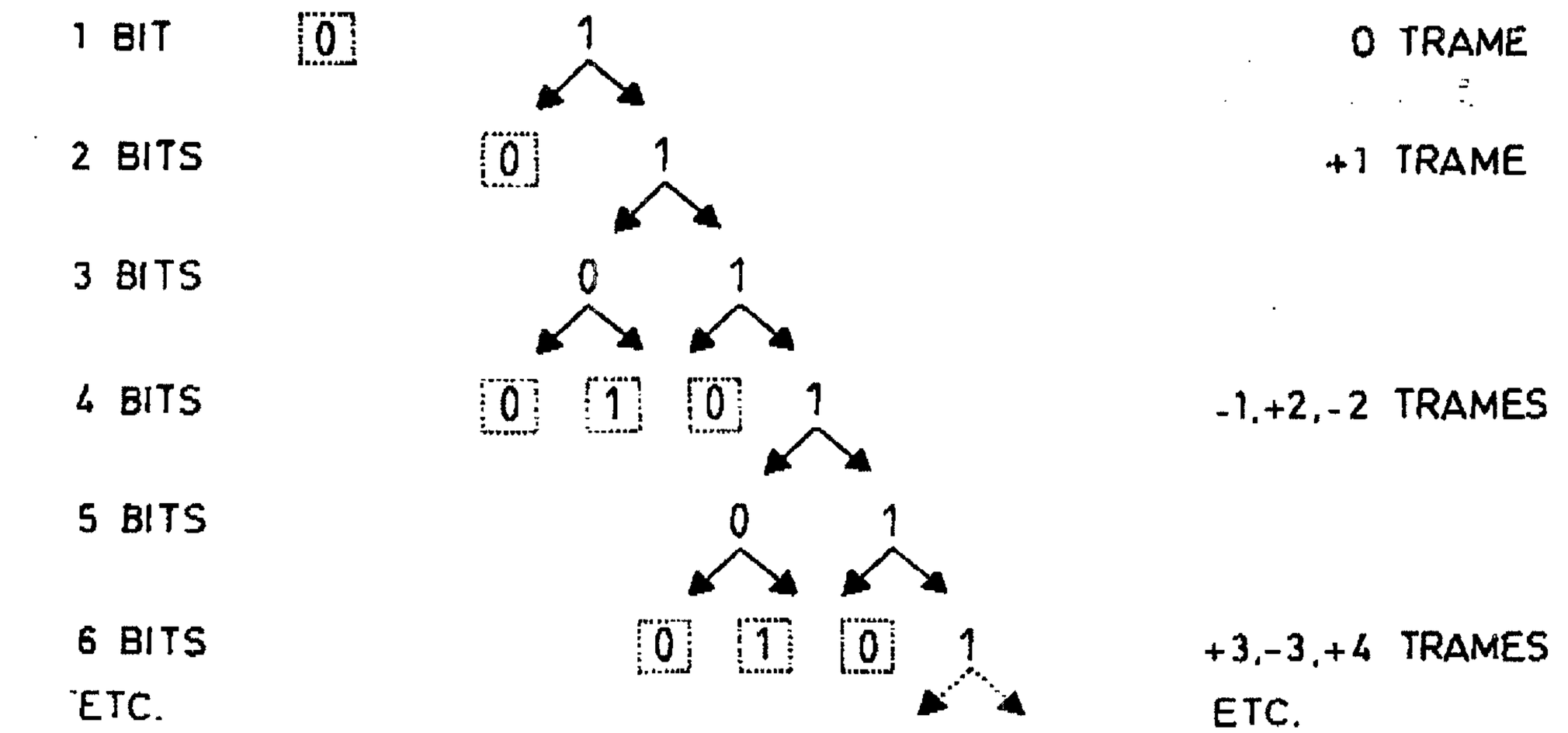

LONGUEUR DU MOT DE CODE
LONGUEUR DU SEGMENT
1 BIT
0
1
3 TRAMES
2 BITS
0
1
4 TRAMES
3 BITS
0
1
4 BITS
0
1
0
1
5,6,7 TRAMES
5 BITS
0
1
6 BITS
0
1
0
1
8,9,10 TRAMES
ETC.
ETC.

FIG.3


LONGUEUR DU MOT DE CODE
DIFFÉRENCE DE LONGUEUR
1 BIT
0
1
0 TRAME
2 BITS
0
1
+1 TRAME
3 BITS
0
1
4 BITS
0
1
0
1
-1,+2,-2 TRAMES
5 BITS
0
1
6 BITS
0
1
0
1
+3,-3,+4 TRAMES
ETC.
ETC.

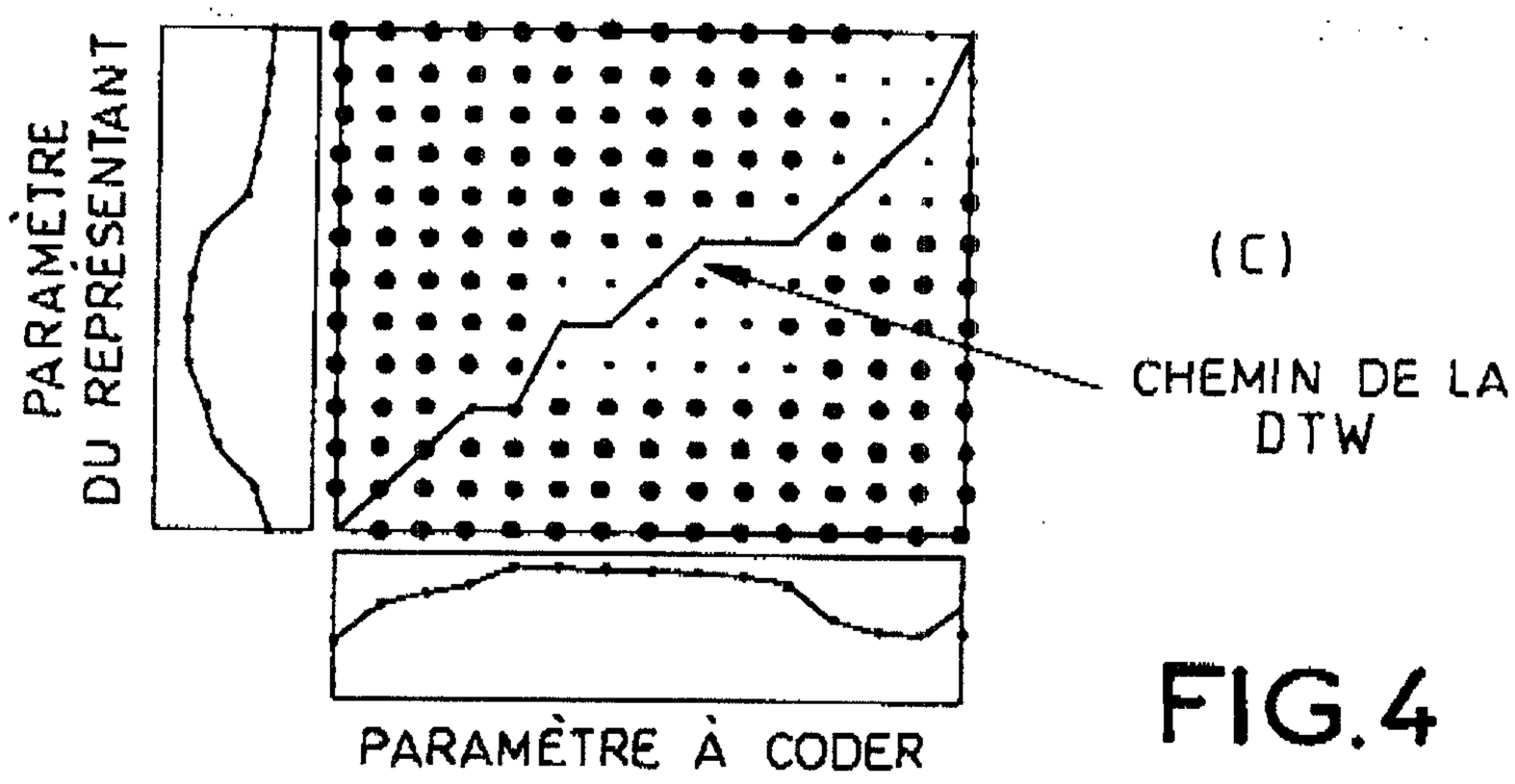
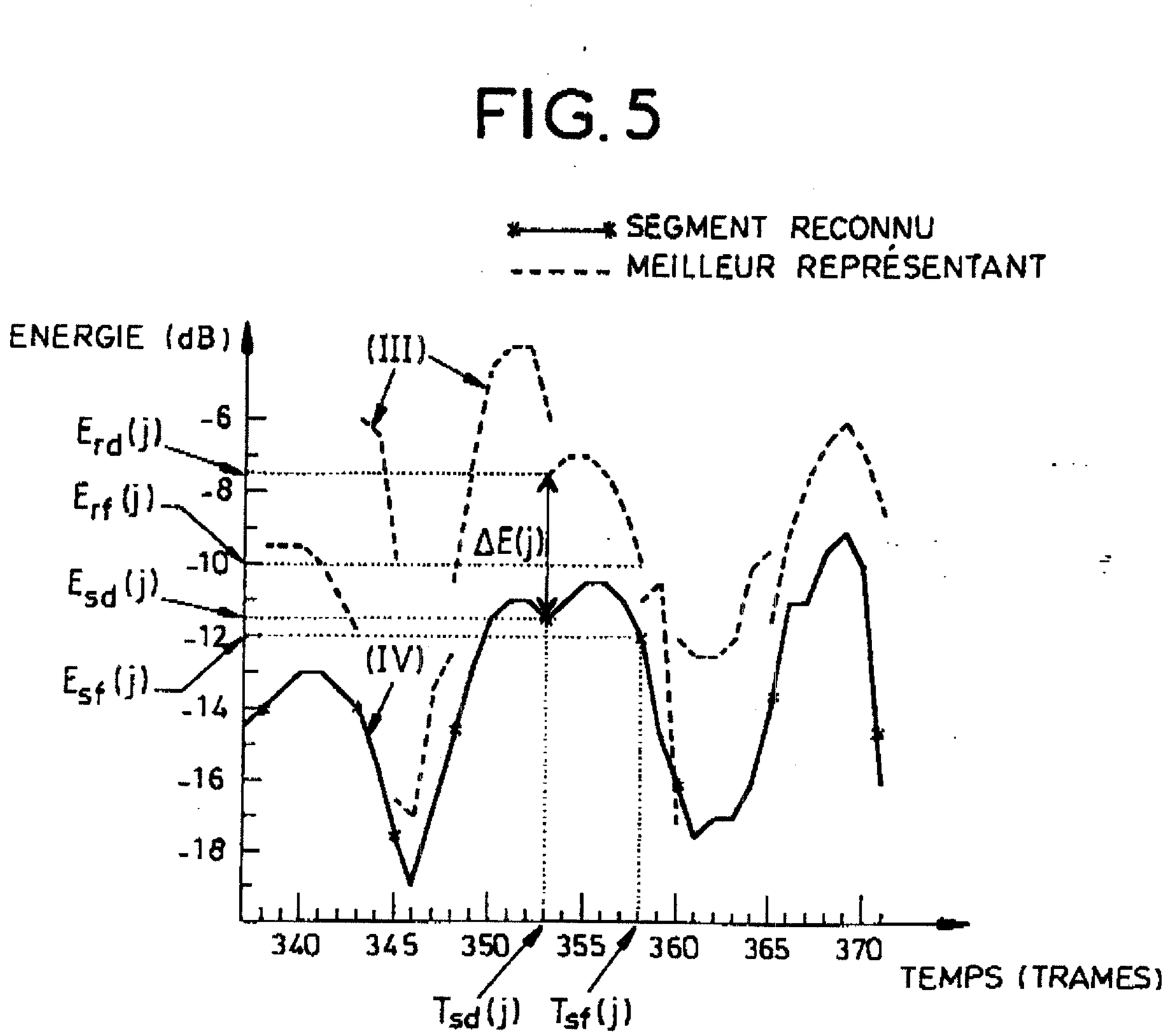

PARAMÈTRE DU REPRÉSENTANT
(C)
CHEMIN DE LA DTW
PARAMÈTRE À CODER

FIG.4

FIG.5


SEGMENT RECONNU
MEILLEUR REPRÉSENTANT
ENERGIE (dB)
(III)
(IV)
$E_{rd}(j)$
$E_{rf}(j)$
$E_{sd}(j)$
$E_{sf}(j)$
ΔE(j)
-6
-8
-10
-12
-14
-16
-18
340
345
350
355
360
365
370
TEMPS (TRAMES)
$T_{sd}(j)$
$T_{sf}(j)$

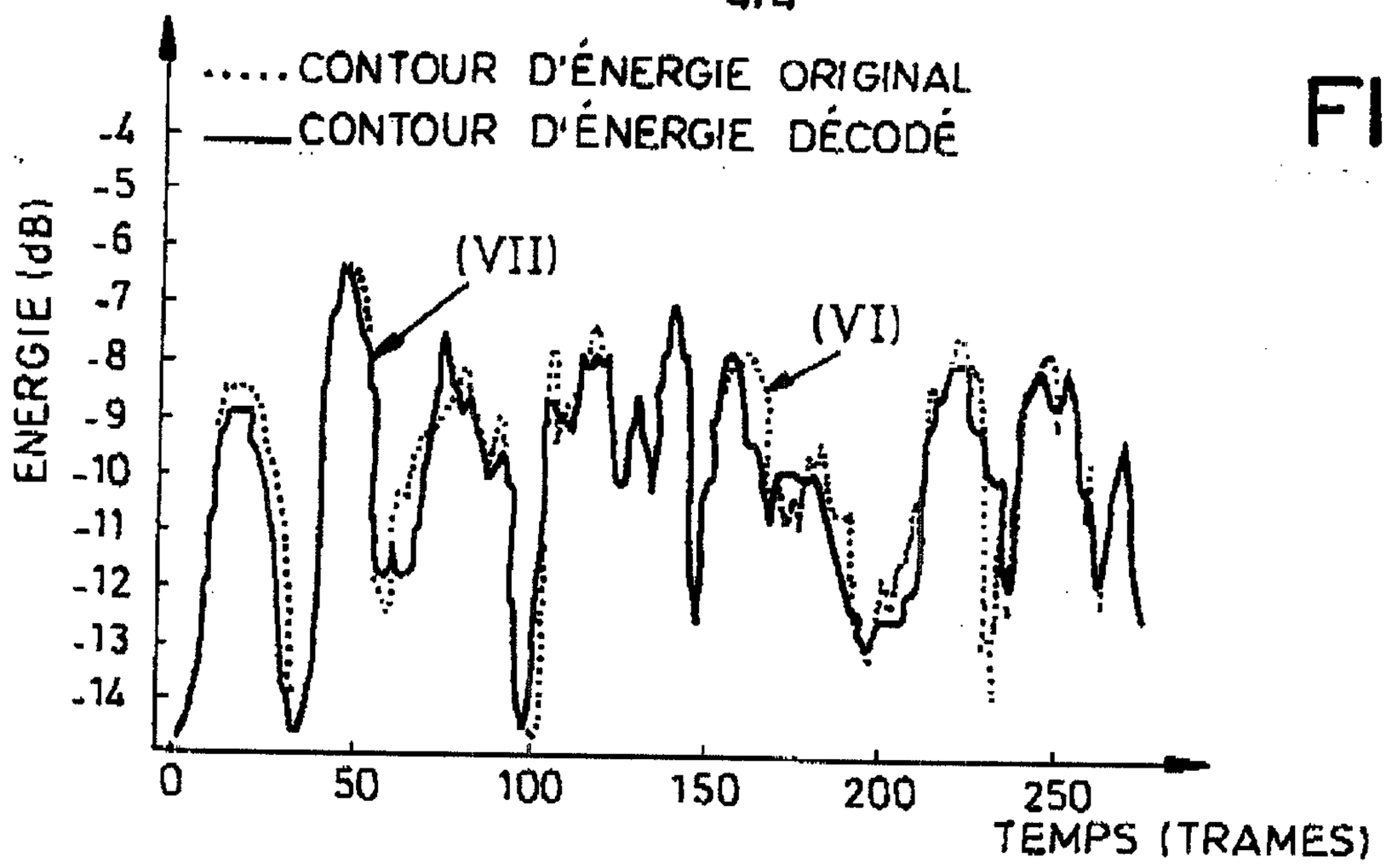


FIG.6

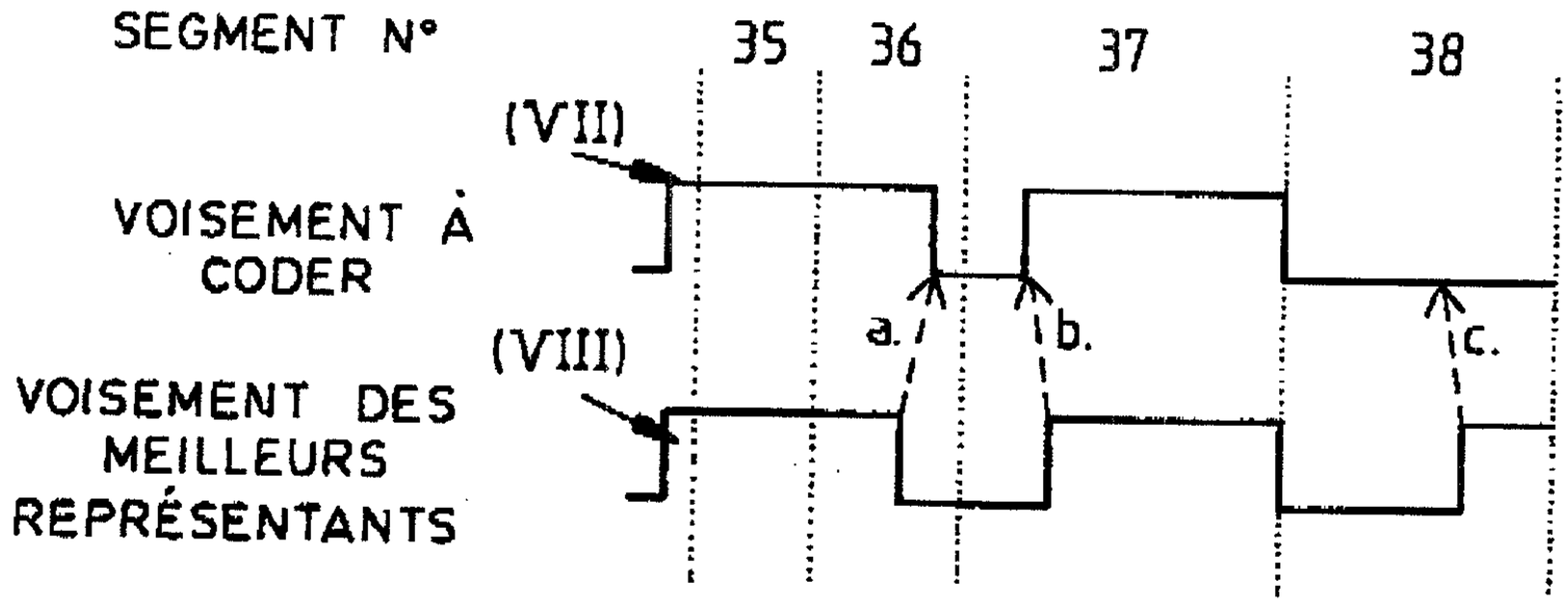


FIG.7

a. : AVANCE D'UNE TRAME
b. : RETARD D'UNE TRAME
c. : TRANSITION À SUPPRIMER

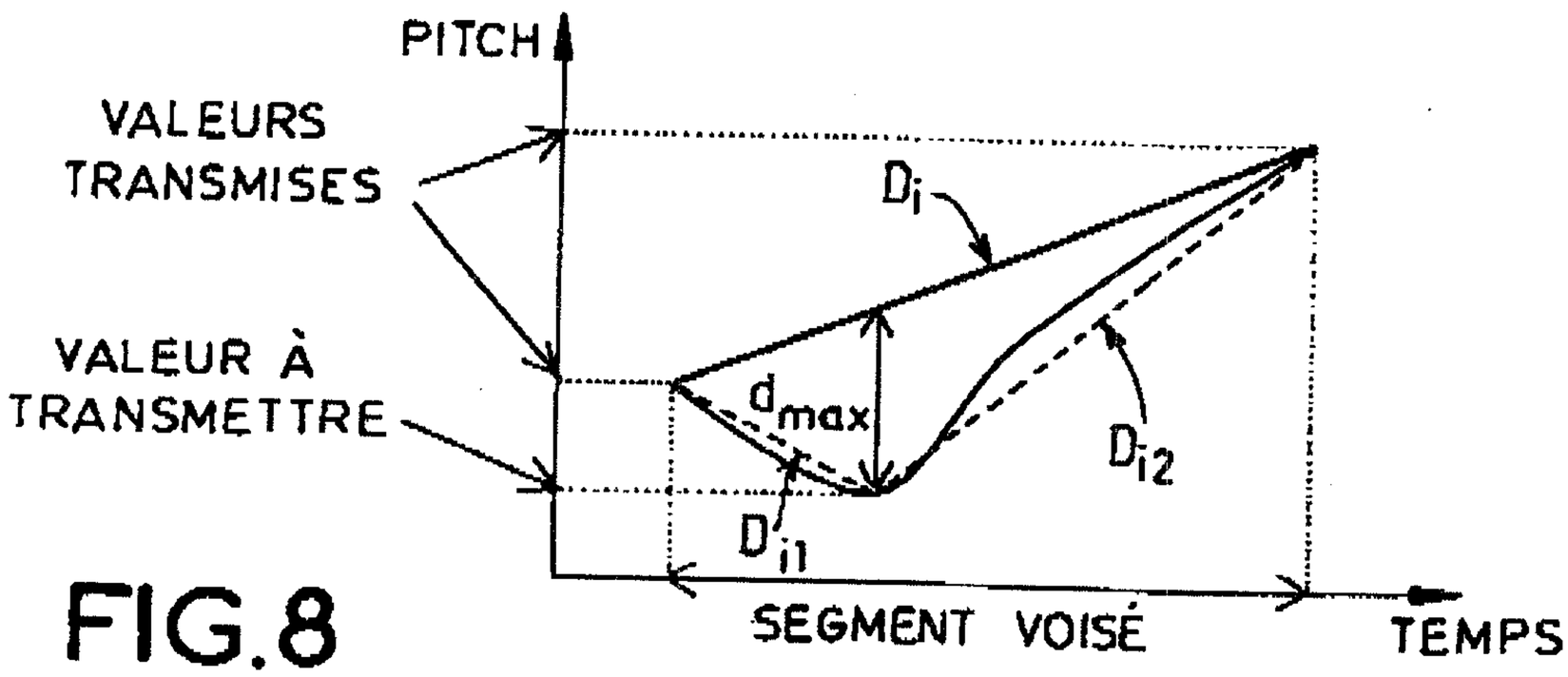
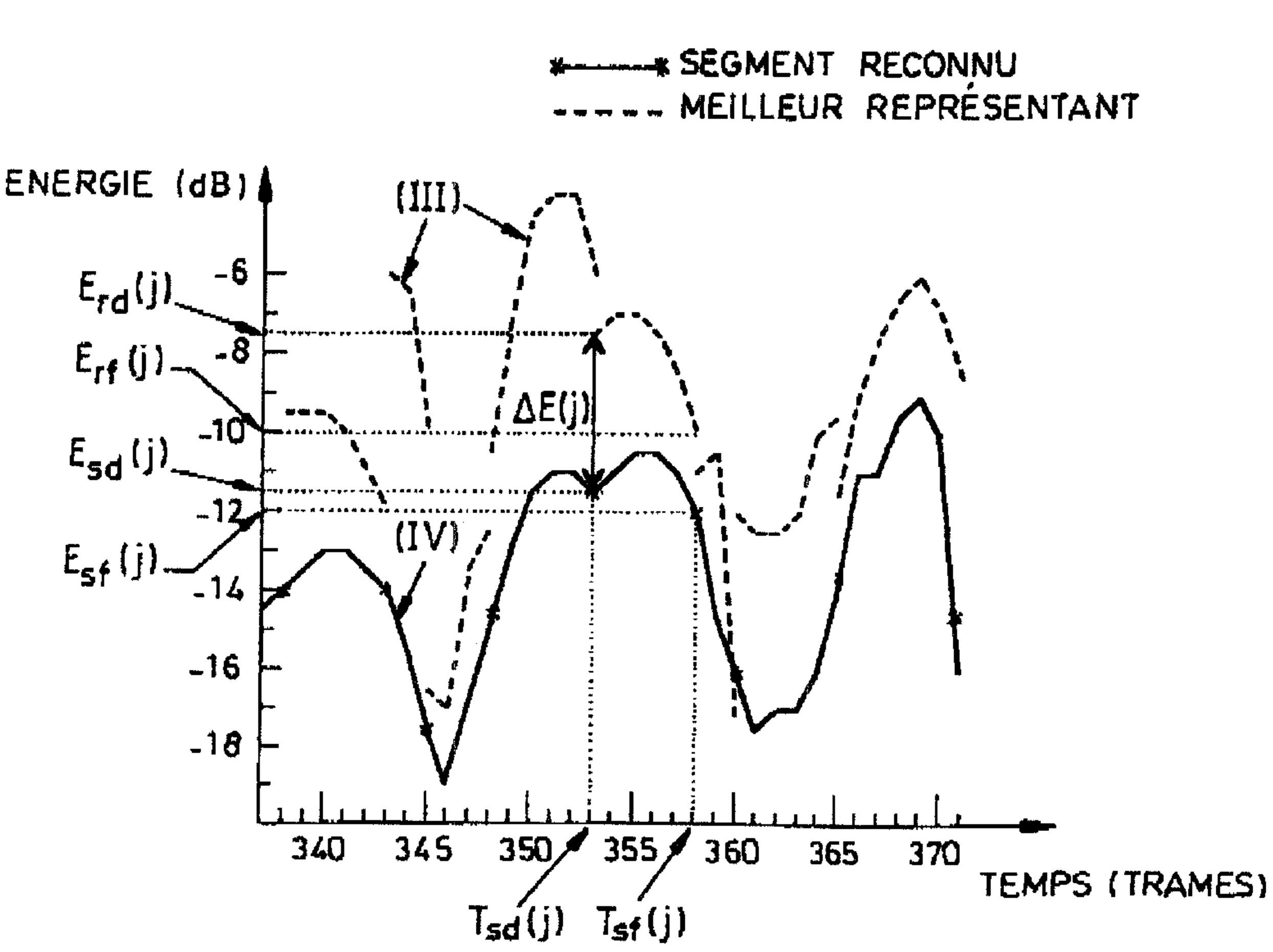


FIG.8


SEGMENT RECONNU
MEILLEUR REPRÉSENTANT
ENERGIE (dB)
-6
-8
-10
-12
-14
-16
-18
$E_{rd}(j)$
$E_{rf}(j)$
$E_{sd}(j)$
$E_{sf}(j)$
(III)
(IV)
$\Delta E(j)$
340
345
350
355
360
365
370
TEMPS (TRAMES)
$T_{sd}(j)$
$T_{sf}(j)$