

①9 RÉPUBLIQUE FRANÇAISE
INSTITUT NATIONAL
DE LA PROPRIÉTÉ INDUSTRIELLE
PARIS

①1 N° de publication :
(à n'utiliser que pour les
commandes de reproduction)

2 553 555

②1 N° d'enregistrement national :

83 16392

⑤1 Int Cl⁴ : G 10 L 5/02.

①2

DEMANDE DE BREVET D'INVENTION

A1

②2 Date de dépôt : 14 octobre 1983.

③0 Priorité :

④3 Date de la mise à disposition du public de la
demande : BOPI « Brevets » n° 16 du 19 avril 1985.

⑥0 Références à d'autres documents nationaux appa-
rentés :

⑦1 Demandeur(s) : Société dite : TEXAS INSTRUMENTS
FRANCE. — FR.

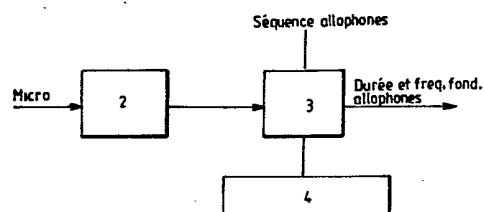
⑦2 Inventeur(s) : Gérard Victor Benbassat.

⑦3 Titulaire(s) :

⑦4 Mandataire(s) : Cabinet Lavoix.

⑤4 Procédé de codage de la parole et dispositif pour sa mise en œuvre.

⑤7 Dispositif de codage de la parole, caractérisé en ce qu'il
comporte des moyens 2 pour analyser et coder la version
parlée du message à coder, et des moyens 3 pour combiner
les codes du message écrit correspondant aux codes du
message parlé et pour engendrer un code de combinaison
contenant des données de durée et de fréquence fondamen-
tale des allophones du message codé.



FR 2 553 555 - A1

D

La présente invention est relative au codage de la parole.

Dans de nombreuses applications, le signal de la parole est codé de telle sorte qu'il peut être
5 stocké de façon numérique en vue d'être émis plus tard ou bien être reproduit localement par un dispositif quelconque.

Dans les deux cas précités, une cadence de bits très faible peut être nécessaire soit pour satis-
10 faire à une exigence de canal de transmission, soit pour permettre le stockage d'un vocabulaire très étendu.

Une faible cadence de bits peut être obtenue en utilisant la synthèse de la parole à partir d'un
15 texte.

Le code obtenu peut être une représentation orthographique du texte lui-même, ce qui permet d'obtenir une cadence de bits de 50 bits/seconde.

Pour simplifier le décodeur utilisé dans une
20 installation de traitement des informations codées de ce type, le code peut être composé d'une séquence de codes de phonèmes et de repères prosodiques obtenus à partir du texte, une telle conception entraînant un léger accroissement de la cadence de bits.

25 Malheureusement, la parole reproduite de cette manière pêche par un important manque de naturel et dans les meilleurs des cas, elle est d'une grande monotonie.

La cause essentielle de ce défaut est l'intonation " synthétique " que l'on obtient avec un tel
30 procédé.

Un tel résultat est très compréhensible lorsque l'on considère la complexité des phénomènes d'intonation qui doivent non seulement satisfaire

certaines règles linguistiques, mais également refléter certains aspects de la personnalité et de l'état d'esprit de la personne qui parle.

5 Il est à l'heure actuelle difficile de prévoir quand des règles prosodiques capables de donner au langage des intonations "humaines" seront disponibles pour tous les langages.

10 Il existe par ailleurs des procédés de codage qui impliquent des cadences de bits beaucoup plus élevées.

15 De tels procédés donnent des résultats satisfaisants mais présentent l'inconvénient essentiel de nécessiter de disposer de mémoires dont la capacité est telle qu'elle rend leur emploi souvent prohibitif.

20 L'invention vise à remédier aux inconvénients précités en créant un procédé de synthèse de la parole qui, tout en nécessitant une cadence de bits relativement peu importante, assure la reproduction de la parole avec des intonations se rapprochant considérablement des intonations naturelles de la voix humaine.

25 Elle a donc pour objet un procédé de codage de la parole, consistant à effectuer un codage de la version écrite d'un message à coder, caractérisé en ce qu'il consiste en outre à coder la version parlée du même message et à combiner aux codes du message écrit les codes des paramètres d'intonation tirés du message parlé.

30 L'invention sera mieux comprise à l'aide de la description qui va suivre, donnée uniquement à titre d'exemple et faite en se référant aux dessins annexés, sur lesquels :

- la Fig.1 est un diagramme montrant le

trajet de correspondance optimale entre les versions parlée et synthétique d'un message à coder par le procédé de l'invention;

5 - la Fig.2 est une vue schématique d'un dispositif de codage de la parole mettant en oeuvre le procédé de l'invention;

- la Fig.3 est une vue schématique d'un dispositif de décodage d'un message codé selon le procédé de l'invention.

10 L'utilisation d'un message sous une forme écrite a pour objectif de produire un modèle acoustique du message dans lequel les limites phonétiques sont connues.

15 Ceci peut être obtenu en utilisant l'une des techniques de synthèse de la parole telles que :

- la synthèse par règle dans laquelle chaque segment acoustique correspondant à chaque phonème du message est obtenu en utilisant des règles acoustiques/phonétiques et qui consiste à calculer les paramètres acoustiques du phonème considéré d'après le contexte dans lequel il doit être réalisé.

(- G.Fant & al., O.V.E. II Synthesis, Strategy Proc. of Speech Comm. Seminar, Stockholm 1962,

25 - L.R.Rabiner, Speech Synthesis by Rule : An acoustic Domain Approach. Bell Syst. Tech. J. 47, 17-37, 1968,

- L.R.Rabiner, A model for synthesizing speech by rule IEEE trans. on Audio and Electr. AU 17, pp 7-13, 1969,

30 - D.H. Klatt, Structure of a Phonological Rule component for a Synthesis by Rule Program, IEEE Trans. ASSP-24, 391-398- 1976).

- la synthèse par concaténation d'unités phonétiques emmagasinées, dans un dictionnaire, ces

5 - F. Emerard Synthèse par diphone et traitement de la Prosodie - Thèse 3ème cycle Univ. des Langues et Lettres, Grenoble, 1977).

Les unités phonétiques sont choisies selon des règles plus ou moins sophistiquées en fonction de la nature des unités et de l'entrée écrite.

30 Le codage de la version écrite du message
 étant réalisé par l'un des procédés connus précités,
 on va maintenant procéder au codage du message parlé
 correspondant.

La version parlée du message est d'abord

numérisée et ensuite analysée afin d'obtenir une représentation acoustique du signal de la parole similaire à celle engendrée à partir de la forme écrite du message que l'on appellera version synthétique.

5 Par exemple, les paramètres cpestraux peuvent être obtenus à partir d'une transformation de Fourier ou, d'une façon plus classique, à partir d'une analyse prédictive linéaire (J.D. Markel, A.H.Gray, Linear Prediction of speech - Springer Verlas, Berlin
10 1976).

Ces paramètres peuvent ensuite être stockés sous une forme qui est adaptée pour calculer une distance spectrale entre chaque tronçon de la version parlée et la version synthétique.

15 Par exemple, si la version synthétique du message est obtenue par concaténation, de segments analysés par prédiction linéaire, la version parlée peut être également analysée en utilisant la prédiction linéaire.

20 Les paramètres de prédiction linéaire peuvent être aisément convertis sous la forme de paramètres cpestraux (J.D. Markel, A.H. Gray) et une distance euclidienne entre les deux jeux de coefficients cpestraux forment une bonne mesure de la distance entre les spectres de faible amplitude.
25

La fréquence fondamentale de la version parlée peut être obtenue en utilisant l'un des nombreux algorithmes de détermination de la fondamentale de signal de parole existants (L.R.Rabiner &
30 al - A comparative performance study of several pitch detection algorithms, IEEE Trans. Acoust. Speech and signal Process, Vol. ASSP 24, pp 399-417 Oct. 1976. B.Secrest, G.Boddington, Post-processing techniques for voice pitch trackers - Procs of the ICASSP 1982 -

Paris, pp. 172-175).

Les versions parlée et synthétique sont ensuite comparées en utilisant une technique de programmation dynamique agissant sur les distances spectrales d'une manière devenue classique dans la reconnaissance globale de la parole (H.SAKOE ET S.CHIBA - Dynamic programming algorithm optimisation for spoken word recognition IEEE trans ASSP-26-1, Fev. 1978).

Cette technique est également appelée compression-extension temporelle dynamique car elle fournit une correspondance élément par élément (ou projection) entre les deux versions du message de sorte que la distance spectrale totale entre elles est réduite au minimum.

Sur la Fig.1, on a représenté en abscisses des unités phonétiques de la version synthétique d'un message et en ordonnées, la version parlée de ce même message dont les segments correspondent respectivement aux unités phonétiques de la version synthétique.

Afin de faire correspondre la durée de la version synthétique avec celle de la version parlée, il est suffisant d'ajuster la durée de chaque unité phonétique de manière à la rendre égale à la durée de chaque segment correspondant de la version parlée.

Après cet ajustement, étant donné que les durées sont égales, la fréquence fondamentale de la version synthétique peut être rendu égale à celle de la version parlée simplement en rendant la fréquence fondamentale de chaque tronçon des unités phonétiques égale à la fréquence fondamentale du tronçon correspondant de la version parlée.

La prosodie est alors composée des compressions-extensions de durée à appliquer à chaque unité

phonétique et du contour de la fréquence fondamentale de la version parlée.

On va maintenant examiner le codage de la prosodie. La prosodie peut être codée de différentes manières qui dépendent du compromis fidélité/cadence de bits souhaité.

Une façon très précise pour effectuer ce codage est la suivante.

Pour chaque tronçon des unités phonétiques, le trajet optimal correspondant peut être vertical, horizontal, ou diagonal.

Si le trajet est vertical, cela signifie que la partie de la version parlée correspondant à ce tronçon est prolongée d'un facteur égal à la longueur du trajet dans un certain nombre de tronçons.

Au contraire, si le trajet est horizontal, tous les tronçons des unités phonétiques situés sous cette portion du trajet doivent être raccourcis d'un facteur qui est égal à la longueur du trajet. Si le trajet est diagonal, les tronçons correspondants des unités phonétiques doivent conserver la même longueur.

Avec une contrainte locale appropriée de la compression-extension temporelle, la longueur des trajets horizontaux et verticaux peut être raisonnablement limitée à trois tronçons. Alors, pour chaque tronçon des unités phonétiques, la compression-extension de durée peut être codée à l'aide de trois bits.

La fréquence fondamentale de chaque tronçon de la version parlée peut être copiée dans chaque tronçon correspondant des unités phonétiques, en utilisant une interpolation d'ordre zéro ou un.

Les valeurs de la fréquence fondamentale peuvent être codées efficacement avec six bits.

Il en résulte qu'un tel codage conduit à un

taux de 9 bits/tronçon pour la prosodie.

Si l'on suppose une moyenne de 40 tronçons/s ceci donne une cadence de l'ordre de 400 bits/s y compris le code phonétique.

5 Un mode de codage plus compact peut être obtenu en employant un nombre limité de caractères pour coder à la fois la compression-extension en durée et le contour de la fréquence fondamentale.

De telles formes peuvent être identifiées
10 pour des segments contenant plusieurs unités phonétiques.

Un choix approprié pour ces segments est la syllabe. Une définition pratique de la syllabe est la suivante :

15

[(groupe consonantique)] voyelle [(groupe consonantique)]

([] = facultatif)

20 Une syllabe correspond à plusieurs unités phonétiques et ses limites peuvent être déterminées automatiquement à partir de la forme écrite du message. Ensuite, les limites de la syllabe peuvent être identifiées sur la version parlée. Puis, si un groupe
25 de contours de fréquences fondamentales caractéristiques de syllabe a été choisi comme caractères représentatifs, chacun d'eux peut être comparé au contour de fréquence fondamentale réel de la syllabe dans la version parlée et on choisit alors le plus proche du
30 contour de fréquence fondamentale réel.

Par exemple, si l'on a 32 caractères, le code de fréquence fondamentale pour une syllabe occupe 5 bits.

En ce qui concerne la durée, une syllabe

peut être scindée en 3 Segments comme indiqué plus haut.

Le facteur de compression-extension temporelle peut être calculé pour chacune des zones comme
5 expliquée pour le procédé précédemment décrit.

Les groupes de trois facteurs de compression extension peut être limité à un nombre fini en choisissant le plus proche dans un jeu de caractères.

Pour 32 caractères, ceci conduit à nouveau à
10 5 bits par syllabe.

La solution qui vient d'être décrite nécessite environ 10 bits/syllabe pour la prosodie ce qui conduit à un total d'environ 120 bits/s y compris le code phonétique.

15 Sur la Fig.2, on a représenté le schéma d'un dispositif de codage de la parole mettant en oeuvre le procédé suivant l'invention.

L'entrée du dispositif est constituée par la sortie d'un microphone non représenté.

20 Elle est appliquée à l'entrée d'un circuit 2 d'analyse et de codage par prédiction linéaire; la sortie de ce circuit est connecté à l'entrée d'un circuit 3 d'élaboration d'algorithme d'adaptation.

Une autre entrée du circuit 3 est connectée
25 à la sortie d'une mémoire 4 qui constitue un dictionnaire d'allophones.

Enfin, sur une troisième entrée 5, le circuit 3 d'élaboration d'algorithme d'adaptation reçoit les séquences d'allophones. Le circuit 3 délivre à sa
30 sortie un message codé contenant la durée et les fréquences fondamentales des allophones.

Afin d'assigner la prosodie d'une phrase à une chaîne d'allophones, la phrase est enregistrée et analysée dans le circuit 3 en utilisant le codage par

prédiction linéaire.

Les allophones sont ensuite comparés avec la phrase codée par prédiction linéaire dans le circuit 3 et des informations de prosodie telles que la durée des allophones et la fréquence fondamentale sont tirées de la phrase et assignées à la chaîne d'allophones.

La cadence des données provenant du microphone à l'entrée du circuit de la Fig.2 étant par exemple de 96.000 bits/s, le message codé correspondant disponible à la sortie de ce circuit a une cadence de 120 bits/s.

La répartition des bits est la suivante.

- 5 bits pour la désignation d'un allophone/phonème (32 valeurs)
- 3 bits pour la durée (7 valeurs)
- 5 bits pour la fréquence fondamentale (7 valeurs)

Ceci fait au total 13 bits par phonème.

Si l'on considère qu'il y a de l'ordre de 9 à 10 phonèmes par seconde, on obtient une cadence de l'ordre de 120 bits/s.

Le circuit représenté à la Fig.3 est le circuit de décodage des signaux engendrés par le circuit de la Fig.2.

Ce dispositif comporte un circuit 6 d'élaboration d'algorithmes de concaténation dont une entrée est destinée à recevoir le message codé à 120 bits/s.

Par une autre entrée, le circuit 6 est connecté à un dictionnaire d'allophones 7. La sortie du circuit 6 est reliée à l'entrée d'un synthétiseur 8, par exemple du type TMS 5200 A. La sortie du synthétiseur 8 est connectée à un haut-parleur 9.

Le circuit 6 délivre un message codé par prédiction linéaire dont la cadence est de 1800 bits/s et le synthétiseur 8 convertit à son tour ce message en un message dont la cadence est de 64000 bits/s
5 exploitable par le haut-parleur 9.

Pour la langue américaine, on a développé un dictionnaire d'allophone comprenant 128 allophones d'une longueur comprise entre 2 et 15 tronçons, la longueur moyenne étant de 4,5 tronçons.

10 Pour la langue française, le processus de concaténation des allophones est différent en ce que le dictionnaire comporte 250 états stables et autant de transitions.

15 Les zones d'interpolation sont utilisées pour rendre plus régulières les transitions entre les allophones du dictionnaire américain.

Les zones d'interpolation sont également utilisées pour régulariser l'énergie au début et à la fin des phrases. Pour obtenir une cadence de données
20 de 120 bits/s, trois bits par phonème sont réservés à l'information de durée.

Le code de durée est le rapport du nombre de tronçons dans l'allophone modifié au nombre de tronçons dans l'original. Ce rapport de codage est
25 nécessaire pour les allophones de la langue américaine car leur longueur peut varier de 1 à 15 tronçons.

En revanche étant donné que les ensembles transitions + états stables de la langue française ont une longueur de 4 à 5 tronçons, leur longueur modifiée
30 peut être égale à 2 à 9 tronçons et le code de durée peut être le nombre de tronçons dans l'ensemble états stables + transitions modifié.

L'invention qui vient d'être décrite permet un codage de la parole avec une cadence de données

relativement faible par rapport aux cadences obtenues par les procédés classiques.

Elle donc particulièrement applicable pour la réalisation de livres dont les pages comportent en
5 parallèle avec des lignes d'écriture ou des images, un texte correspondant codé et reproductible par un synthétiseur.

Elle est également très intéressante à utiliser dans les systèmes de vidéo-texte développés par
10 la Demanderesse et notamment dans les dispositifs pour l'audition de messages parlés synthétisés et pour la visualisation de messages graphiques correspondants du type décrit à la demande de brevet FR. 83 09 194 déposée le 2 Juin 1983 par la Demanderesse.

REVENDICATIONS

1. Procédé de codage de la parole, consistant à effectuer un codage de la version écrite d'un message à coder, caractérisé en ce qu'il consiste
5 en outre à coder la version parlée du même message et à combiner aux codes du message écrit, les codes des paramètres d'intonation tirés du message parlé.

2. Procédé suivant la revendication 1, caractérisé en ce que la version écrite est utilisée
10 pour engendrer les composants en segments du message.

3. Procédé suivant l'une quelconque des revendications 1 et 2, caractérisé en ce que la version parlée du message à coder est analysée puis comparée avec les segments concaténés obtenus à partir
15 de la version écrite afin de déterminer l'alignement correct dans le temps entre les deux versions.

4. Procédé suivant la revendication 3, caractérisé en ce que les composants de la forme écrite étant engendrés par concaténation de petits
20 segments sonores stockés dans un dictionnaire, la version parlée est comparée avec lesdits segments concaténés en utilisant un algorithme de programmation dynamique.

5. Procédé suivant la revendication 4, caractérisé en ce que ladite programmation dynamique
25 opère sur les distances spectrales.

6. Dispositif de codage de la parole destiné à mettre en oeuvre le procédé suivant l'une quelconque des revendications 1 à 5, caractérisé en ce qu'il com-
30 porte des moyens (2) pour analyser et coder la version parlée du message à coder, et des moyens (3) pour combiner les codes du message écrit correspondant aux codes du message parlé et pour engendrer un code de combinaison contenant des données de durée et de fré-

quence fondamentale des allophones du message codé.

7. Dispositif suivant la revendication 6, caractérisé en ce que lesdits moyens d'analyse et de codage de la version parlée du message à coder sont constitués par un circuit d'analyse et de codage par prédiction linéaire.

8. Dispositif suivant l'une des revendications 5 à 7, caractérisé en ce que lesdits moyens (3) de combinaison des codes de la version parlée à ceux de la version écrite du message à coder comprennent des moyens pour élaborer un algorithme d'adaptation auxquels est associé un dictionnaire d'allophones (4) en vue de la synthèse par concaténation des composants de la version écrite.

9. Dispositif de décodage d'un message codé par le procédé suivant l'une quelconque des revendications 1 à 5, caractérisé en ce qu'il comporte des moyens (6) d'élaboration d'un algorithme de concaténation en vue d'engendrer des signaux codés par prédiction linéaire à partir du code résultant de la combinaison des codes de la version écrite et de la version parlée du message et de données contenues dans un dictionnaire d'allophones (7) associé, et un synthétiseur de la parole (8) associé à des moyens (9) de reproduction sonore.

