(19) RÉPUBLIQUE FRANÇAISE

INSTITUT NATIONAL
DE LA PROPRIÉTÉ INDUSTRIELLE

PARIS

(11) **N° de publication :** **2 802 690**
(à n'utiliser que pour les commandes de reproduction)

(21) **N° d'enregistrement national :** **99 15953**

(51) Int Cl[7] : **G 08 C 23/02,** G 10 L 15/14

(12) DEMANDE DE BREVET D'INVENTION A1

(22) **Date de dépôt :** 17.12.99.

(30) **Priorité :**

(43) **Date de mise à la disposition du public de la demande :** 22.06.01 Bulletin 01/25.

(56) **Liste des documents cités dans le rapport de recherche préliminaire :** *Se reporter à la fin du présent fascicule*

(60) **Références à d'autres documents nationaux apparentés :**

(71) **Demandeur(s) :** *THOMSON MULTIMEDIA Société anonyme* — FR.

(72) **Inventeur(s) :** SOUFFLET FREDERIC et JOUET PIERRICK.

(73) **Titulaire(s) :**

(74) **Mandataire(s) :** THOMSON MULTIMEDIA.

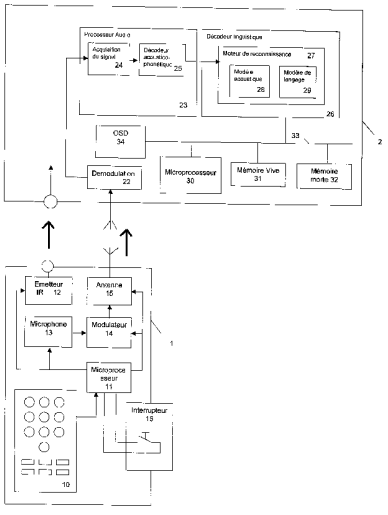
(54) PROCEDE ET DISPOSITIF DE RECONNAISSANCE VOCALE, DISPOSITIF DE TELECOMMANDE ASSOCIE.

(57) L'invention concerne un dispositif de reconnaissance vocale.

Selon l'invention, le dispositif comporte

- un circuit (23, 24, 25) d'acquisition d'un signal comportant des données vocales en provenance d'un utilisateur,
- des moyens de détection (22, 30) d'un signal de fin de données vocales généré par intervention de l'utilisateur,
- des moyens d'analyse (26) des données vocales aptes à modifier l'évolution de l'analyse en fonction du signal de fin de données vocales.

L'invention concerne également un dispositif de télécommande pour déclencher le signal de fin de données vocales, ainsi qu'un procédé.



FR 2 802 690 - A1

L'invention concerne un dispositif de reconnaissance vocale à déclenchement volontaire de certaines phases de la reconnaissance. L'invention concerne également un dispositif pour réaliser le déclenchement, en particulier à distance. L'invention s'applique notamment dans le domaine de la télévision.

Un système de reconnaissance vocale typique comporte d'une part un processeur audio incluant des moyens d'acquisition et de traitement d'un signal audio représentatif des données vocales à reconnaître et d'autre part un décodeur linguistique comportant le moteur de reconnaissance vocale proprement dit. Ce moteur utilise un modèle acoustique et un modèle de langage pour réaliser la reconnaissance sur la base des signaux audio prétraités par le processeur audio.

En particulier lorsque le modèle de langage est basé sur des grammaires, l'analyse d'une phrase par le moteur de reconnaissance ne commence qu'après l'expiration d'un délai prédéterminé durant lequel aucun signal audio n'est reçu. On considère alors que l'interlocuteur du système a fini de prononcer sa phrase.

Selon l'application envisagée, le choix du délai devient cornélien. S'il est choisi trop long, le retard de traitement d'une phrase peut devenir rédhibitoire. S'il est choisi trop court, alors des hésitations dans l'énonciation de la phrase par l'utilisateur peuvent déclencher le traitement avant que cette énonciation ne soit terminée. De telles hésitations apparaissent par exemple lorsque l'interlocuteur prend connaissance, en même temps qu'il commence sa phrase, de données s'affichant sur un écran en réponse à des actions précédentes.

Pour éviter les déclenchements de traitement intempestifs suite à des hésitations, on peut envisager d'allonger le délai prédéterminé, dont la durée peut aller au-delà de cinq ou six secondes. Dans l'application envisagée ici, en l'occurrence la commande vocale d'un récepteur de télévision et d'applications s'y rapportant, cet ordre de grandeur de délai est incompatible avec les attentes du consommateur.

L'invention a pour objet un dispositif de reconnaissance vocale caractérisé en ce qu'il comporte

- un circuit d'acquisition d'un signal comportant des données vocales en provenance d'un utilisateur,
- des moyens de détection d'un signal de fin de données vocales généré par intervention de l'utilisateur,
- des moyens d'analyse des données vocales aptes à modifier l'évolution de l'analyse en fonction du signal de fin de données vocales.

Ainsi, l'utilisateur peut intervenir directement sur l'analyse, en signifiant qu'il a fini d'énoncer son texte.

Selon un mode de réalisation particulier, les moyens d'analyse des données vocales finalisent l'analyse des données vocales précédemment stockées sur réception du signal de fin de données vocales.

Selon un mode de réalisation particulier, les moyens d'analyse mettent en œuvre un algorithme de type Viterbi et le retraçage à travers des états passés pour déterminer une ou plusieurs séquences de mots susceptibles de correspondre aux données vocales est démarré dès réception du signal de fin de données vocales.

Selon un mode de réalisation particulier, le signal de fin de données est généré par activation manuelle d'un moyen de génération de signal par l'utilisateur.

Selon un mode de réalisation particulier, le moyen de génération de signal de fin de données comporte un interrupteur d'une télécommande.

Selon un mode de réalisation particulier, le signal comportant les données vocales est reçu par transmission sans fil.

L'invention a aussi pour objet un dispositif de télécommande comportant un microphone pour générer un signal comportant des données vocales et des circuits d'émission du signal comportant des données vocales caractérisé en ce qu'il comporte en outre des moyens de génération et

d'émission d'un signal de fin de données vocales actionnables par un utilisateur.

Selon un mode de réalisation particulier, les moyens de génération de signal de fin de données vocales comprennent un interrupteur actionnable par l'utilisateur.

Selon un mode de réalisation particulier, l'interrupteur est disposé de manière à contrôler le fonctionnement des circuits d'émission du signal comportant des données vocales.

Selon un mode de réalisation particulier, le signal de fin de données vocales est constitué par le passage de la présence de porteuse du signal comportant des données vocales à l'absence de porteuse.

L'invention a aussi pour objet un procédé de reconnaissance vocale caractérisé en ce qu'il comporte les étapes :

- d'acquisition d'un signal comportant des données vocales,
- d'analyse du signal acquis en vue de la recherche de mots ou de séquences de mots représentatifs du signal acquis, l'analyse comportant plusieurs phases successives,
- de conditionnement de franchissement d'au moins une phase à la réception d'un signal de fin de données vocales déclenché par un utilisateur.

Selon un mode de réalisation particulier, l'étape d'analyse du signal acquis comporte une phase de détermination en parallèle d'une pluralité de mots ou de séquences de mots candidats représentatifs du signal acquis, et une phase de choix d'un mot ou d'une séquence de mots parmi candidats.

D'autres caractéristiques et avantages de l'invention apparaîtront à travers la description d'un exemple de réalisation particulier non limitatif. Cet exemple sera décrit en liaison avec les dessins joints parmi lesquels

- la figure 1 est un diagramme d'un système de réception de télévision mettant en œuvre un sous-système de reconnaissance vocale,
- la figure 2 est un organigramme d'un exemple de mise en œuvre du procédé objet de l'invention.

Le système de la figure 1 comprend une télécommande 1 et un récepteur de télévision 2.

La télécommande 1 comporte de façon connue un clavier de touches 10, un microprocesseur 11 configuré pour recevoir les signaux en provenance du clavier 10, et un circuit de modulation analogique et de transmission par ondes infrarouges 12, pour émission vers le téléviseur 2.

La télécommande 1 comporte en outre un microphone 13 relié à un circuit de modulation en radio fréquences 14. Ce circuit 14 est relié à une antenne 15, pour émission des signaux RF vers le téléviseur 2. Le circuit de modulation 14 et le microphone 13 sont contrôlés par le microprocesseur.

La télécommande est également équipée d'un interrupteur 16, relié au microprocesseur 11.

La voie infrarouges de la télécommande fonctionne de façon classique. La voie radiofréquences fonctionne de la manière suivante : lorsque l'utilisateur actionne l'interrupteur 16, le microprocesseur 11 commande de façon appropriée le circuit de modulation et le microphone pour que les signaux vocaux de l'utilisateur soient traités et transmis par l'antenne 15. Lorsque l'interrupteur n'est pas actionné, l'alimentation de l'ensemble des organes nécessaires à la voie radiofréquences est coupée, dans le but de réduire leur consommation.

Un signal RF n'est donc transmis au téléviseur que lorsque l'interrupteur est actionné.

Une télécommande d'un type similaire est décrite dans la demande de brevet français FR 9804847, déposée le 17 avril 1998 au nom de THOMSON multimedia et publiée le 22 octobre 1999 sous le numéro FR 2777681.

Le rôle de la télécommande est donc simplement d'acquérir le signal audio et de le transmettre sous forme analogique au téléviseur. Dans le cadre du présent exemple, le traitement effectué par la télécommande est réduit au minimum pour limiter sa consommation électrique.

Le récepteur de télévision 2 comporte une antenne 20 pour la réception des signaux en provenance de l'antenne de la télécommande, ainsi qu'un circuit de réception infrarouge 21. L'antenne 20 est reliée à un circuit de syntonisation et de démodulation 22. Le signal démodulé est transmis à un

processeur audio 23 comportant un circuit d'acquisition 24 et un décodeur acoustico-phonétique 25. Le circuit d'acquisition est muni d'un convertisseur analogique numérique (non illustré) pour réaliser l'échantillonnage du signal audio en bande de base à une fréquence de 22 kHz.

Le décodeur acoustico-phonétique traduit les échantillons numériques en symboles acoustiques choisis dans un alphabet prédéterminé.

Un décodeur linguistique 26 traite ces symboles dans le but de déterminer, pour une séquence A de symboles, la séquence W de mots la plus probable, étant donné la séquence A. Le décodeur linguistique 26 comporte un moteur de reconnaissance 27 utilisant un modèle acoustique 28 et un modèle de langage 29. Le modèle acoustique est par exemple un modèle dit 'Markov caché ('Hidden Markov Model' ou HMM). Il calcule de façon connue en soi les scores acoustiques des séquences de mots considérées. Le modèle de langage mis en œuvre dans le présent exemple de réalisation est basé sur une grammaire décrite à l'aide de règles de syntaxe de forme Backus Naur. Le modèle de langage est utilisé pour déterminer une pluralité d'hypothèses de séquences de mots et pour calculer des scores linguistiques.

Le moteur de reconnaissance est basé sur un algorithme de type Viterbi appelé 'n-best'. L'algorithme de type n-best détermine à chaque étape de l'analyse d'une phrase les n séquences de mots les plus probables. En fin de phrase, la solution la plus probable est choisie parmi les n candidats, à partir des scores fournis par le modèle acoustique et le modèle de langage.

Le récepteur de télévision comprend en outre un microprocesseur 30, une mémoire vive 31 et une mémoire morte 32, connectés à un bus interne 33. Bien que le processeur audio et le décodeur linguistique soient représentés comme des circuits séparés sur la figure 1, au moins le décodeur acoustico-phonétique et le décodeur linguistique peuvent être implémentés sous la forme d'un logiciel stocké dans la mémoire morte 32 et exécuté par le microprocesseur 30.

Le récepteur de télévision comprend également un circuit d'affichage sur écran ('OSD') 34 apte à générer des signaux vidéo représentatifs de menus de commande du récepteur, de textes et/ou de graphiques. Le circuit 34 est également contrôlable par des applications de type guide de programmes

électronique exécutées par le microprocesseur 30. Le cas échéant, les signaux générés par le circuit 34 viennent remplacer partiellement ou totalement ceux issus des circuits de traitement du signal vidéo (non illustrés) reçu par antenne. Un tube cathodique (non illustré) muni des circuits de déflexion appropriés permet d'afficher les signaux vidéo.

On décrira maintenant plus particulièrement le fonctionnement du moteur de reconnaissance. Comme mentionné, ce dernier utilise un algorithme de type Viterbi (algorithme n-best) pour analyser une phrase composée d'une séquence de symboles (vecteurs) acoustiques. L'algorithme détermine les N séquences de mots les plus probables, étant donné la séquence A de symboles acoustiques observée jusqu'au symbole courant. Les séquences de mots les plus probables sont déterminées à travers le modèle de langage de type grammaire stochastiques. En liaison avec les modèles acoustiques des éléments terminaux de la grammaire, qui sont basés sur des HMM (Modèles de Markov cachés ou 'Hidden Markov Models'), on produit alors un modèle de Markov caché global pour l'application, qui inclut donc le modèle de langage et par exemple les phénomènes de coarticulations entre éléments terminaux. L'algorithme de Viterbi est mis en œuvre en parallèle, mais au lieu de retenir une seule transition vers chaque état lors de l'itération i, on retient pour chaque état les N transitions les plus probables.

Des informations concernant en particulier les algorithmes de Viterbi, de recherche par faisceau et de 'n-best' sont données dans l'ouvrage :

"Statistical methods for speech recognition" par Frederick Jelinek, MIT Press 1999 ISBN 0-262-10066-5, chapitres 2 et 5 en particulier.

L'analyse effectuée par le moteur de reconnaissance s'arrête lorsque l'ensemble des symboles acoustiques relatifs à une phrase ont été traités. Le moteur de reconnaissance dispose alors d'un treillis constitué des états à chaque itération précédente de l'algorithme et des transitions entre ces états, jusqu'aux états finaux. En dernier lieu, on retient parmi les états finaux et leurs N transitions associées les N transitions les plus probables. En retraçant les transitions à partir des états finaux, on détermine les N séquences de mots les plus probables correspondant aux symboles acoustiques. Ces séquences sont alors soumises à un traitement utilisant un parseur dans le but de sélectionner l'unique séquence finale sur des critères grammaticaux.

Selon le présent exemple de réalisation, le dernier symbole à analyser avant de procéder au retraçage est supposé reçu une fois que l'interlocuteur relâche l'interrupteur 16 de la télécommande. La télécommande n'émet alors plus de porteuse RF. Cette absence de porteuse est détectée de façon connue par le circuit de syntonisation 22, qui avertit le microprocesseur du récepteur par une interruption appropriée. Le moteur de reconnaissance termine alors son analyse sur la base des symboles acoustiques reçus et fournit la séquence de mots la plus probable à l'application qui gère le guide de programmes.

Ceci permet de prendre en compte un signal volontaire de la part de l'utilisateur de terminer l'analyse de la phrase en cours. Le signal vocal et l'information de fin de phrase ne sont donc pas corrélés.

Selon une variante de réalisation, le récepteur suppose que l'interlocuteur a fini d'énoncer son texte lorsqu'arrive le premier des événements suivants : détection d'absence de porteuse ou détection de silence pendant un intervalle de temps déterminé.

Selon un mode de réalisation particulier, la télécommande émet un signal spécifique suite au relâchement de l'interrupteur 16 et avant de couper l'alimentation du microphone et des circuits d'émission, dans le but de faciliter la détection du relâchement par le récepteur. Ce signal spécifique est par exemple une salve à une fréquence particulière.

Selon un mode de réalisation particulier de l'invention, l'alimentation n'est coupée qu'après une temporisation prédéterminée, dans le but d'éviter les conséquences d'un relâchement provisoire involontaire de l'interrupteur 16. Cette temporisation est par exemple de l'ordre d'une demi-seconde. Si l'interrupteur 16 est de nouveau actionné durant cette temporisation, alors l'alimentation du microphone et des circuits d'émission de la télécommande est maintenue.

Bien que le signal de fin de données vocales soit déclenché grâce à une télécommande dans l'exemple de réalisation décrit ci-dessus, d'autres moyens peuvent être utilisés, notamment des touches du dispositif récepteur.

Revendications

1. Dispositif de reconnaissance vocale caractérisé en ce qu'il comporte

- un circuit (23, 24, 25) d'acquisition d'un signal comportant des données vocales en provenance d'un utilisateur,
- des moyens de détection (22, 30) d'un signal de fin de données vocales généré par intervention de l'utilisateur,
- des moyens d'analyse (26) des données vocales aptes à modifier l'évolution de l'analyse en fonction du signal de fin de données vocales.

2. Dispositif selon la revendication 1, caractérisé en ce que les moyens d'analyse des données vocales finalisent l'analyse des données vocales précédemment stockées sur réception du signal de fin de données vocales.

3. Dispositif selon les revendications 1 ou 2, caractérisé en ce que les moyens d'analyse mettent en œuvre un algorithme de type Viterbi et en ce que le retraçage à travers des états passés pour déterminer une ou plusieurs séquences de mots susceptibles de correspondre aux données vocales est démarré dès réception du signal de fin de données vocales.

4. Dispositif selon l'une des revendications 1 à 3, caractérisé en ce que le signal de fin de données est généré par activation manuelle d'un moyen de génération (16) de signal par l'utilisateur.

5. Dispositif selon la revendication 4, caractérisé en ce que le moyen de génération de signal de fin de données comporte un interrupteur (16) d'une télécommande (1).

6. Dispositif selon l'une des revendications 1 à 5, caractérisé en ce que le signal comportant les données vocales est reçu par transmission sans fil.

7. Dispositif de télécommande (1) comportant un microphone (13) pour générer un signal comportant des données vocales et des circuits d'émission (14, 15) du signal comportant des données vocales caractérisé en

ce qu'il comporte en outre des moyens (11, 14, 15, 16) de génération et d'émission d'un signal de fin de données vocales actionnables par un utilisateur.

8. Dispositif selon la revendication 7, caractérisé en ce que les moyens de génération de signal de fin de données vocales comprennent un interrupteur (16) actionnable par l'utilisateur.

9. Dispositif selon la revendication 8, caractérisé en ce que l'interrupteur (16) est disposé de manière à contrôler le fonctionnement des circuits d'émission (14, 15) du signal comportant des données vocales.

10. Dispositif selon l'une des revendications 7 ou 8, caractérisé en ce que le signal de fin de données vocales est constitué par le passage de la présence de porteuse du signal comportant des données vocales à l'absence de porteuse.

11. Procédé de reconnaissance vocale caractérisé en ce qu'il comporte les étapes :

- d'acquisition d'un signal comportant des données vocales,
- d'analyse du signal acquis en vue de la recherche de mots ou de séquences de mots représentatifs du signal acquis, l'analyse comportant plusieurs phases successives,
- de conditionnement de franchissement d'au moins une phase à la réception d'un signal de fin de données vocales déclenché par un utilisateur.

12. Procédé selon la revendication 11, caractérisé en ce que l'étape d'analyse du signal acquis comporte une phase de détermination en parallèle d'une pluralité de mots ou de séquences de mots candidats représentatifs du signal acquis, et une phase de choix d'un mot ou d'une séquence de mots parmi candidats.

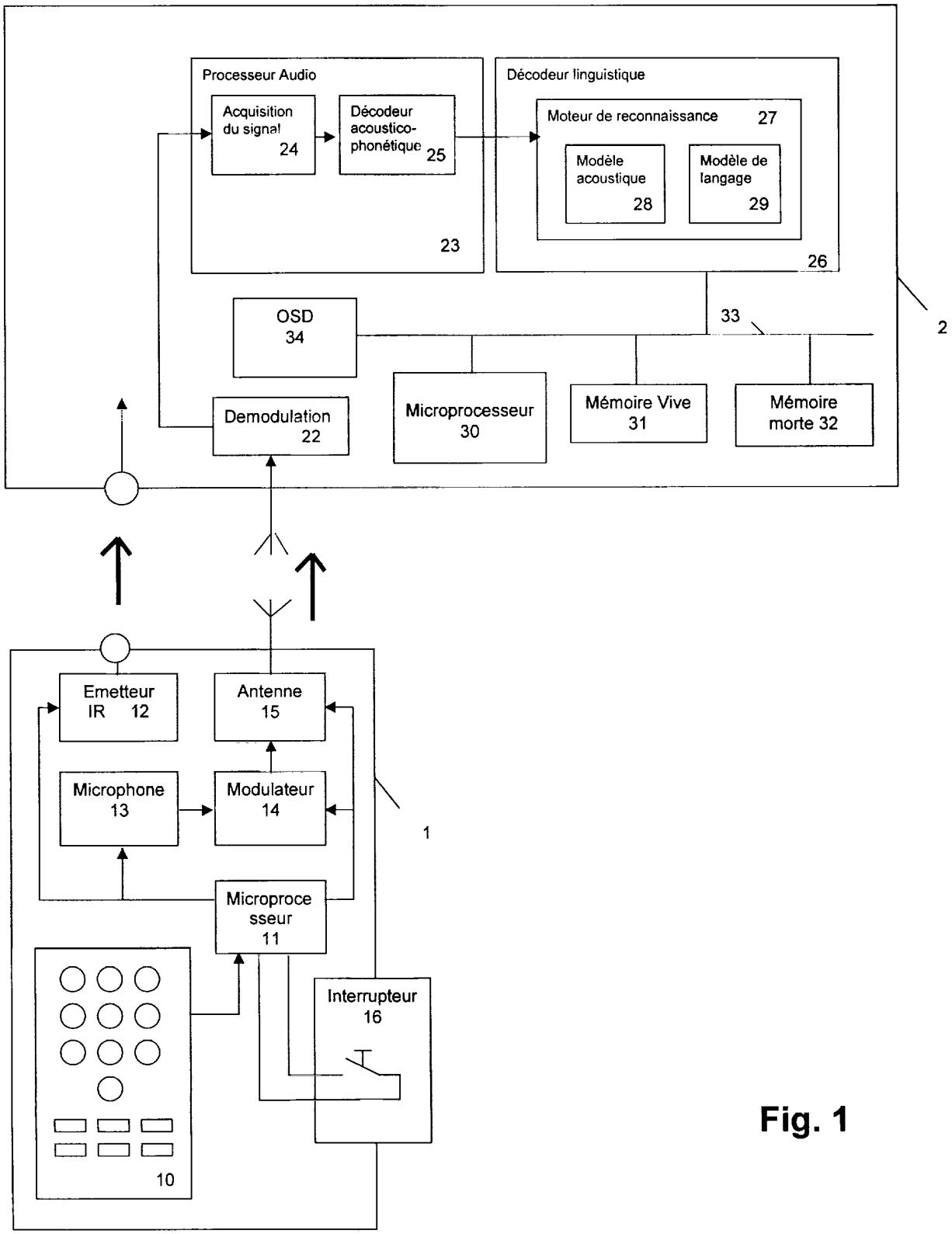

Processeur Audio
Acquisition du signal 24
Décodeur acoustico-phonétique 25
23
Décodeur linguistique
Moteur de reconnaissance 27
Modèle acoustique 28
Modèle de langage 29
26
OSD 34
33
2
Demodulation 22
Microprocesseur 30
Mémoire Vive 31
Mémoire morte 32
Emetteur IR 12
Antenne 15
Microphone 13
Modulateur 14
1
Microprocesseur 11
Interrupteur 16
10

Fig. 1

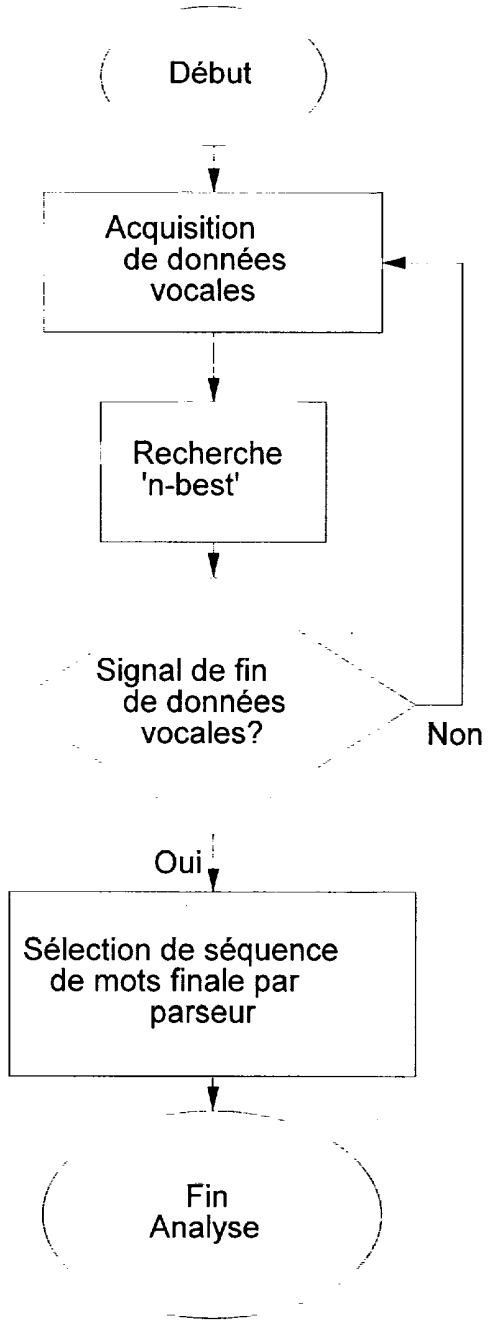

Début
Acquisition de données vocales
Recherche 'n-best'
Signal de fin de données vocales?
Non
Oui
Sélection de séquence de mots finale par parseur
Fin Analyse

Fig. 2

RÉPUBLIQUE FRANÇAISE

INPI
INSTITUT NATIONAL DE LA PROPRIETE INDUSTRIELLE

RAPPORT DE RECHERCHE PRÉLIMINAIRE

établi sur la base des dernières revendications déposées avant le commencement de la recherche

2802690

N° d'enregistrement national

FA 580812
FR 9915953

DOCUMENTS CONSIDÉRÉS COMME PERTINENTS		Revendication(s) concernée(s)	Classement attribué à l'invention par l'INPI
Catégorie	Citation du document avec indication, en cas de besoin, des parties pertinentes		
X	EP 0 078 014 A (NISSAN MOTOR) 4 mai 1983 (1983-05-04) * abrégé; revendication 1; figures 1,4 * ---	1,4,5,7, 8,11	G08C23/02 G10L15/14
A	O'SHAUGHNESSY D: "ANALYSIS AND AUTOMATIC RECOGNITION OF FALSE STARTS IN SPONTANEOUS SPEECH" PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING (ICASSP),US,NEW YORK, IEEE, vol. -, 27 avril 1993 (1993-04-27), pages II-724-727, XP000427892 ISBN: 0-7803-0946-4 * abrégé * ---	1,7,11	
A	EP 0 757 342 A (AT & T CORP) 5 février 1997 (1997-02-05) * colonne 1, ligne 33 - ligne 36 * * colonne 2, ligne 25 - ligne 28 * -----	3	
			DOMAINES TECHNIQUES RECHERCHÉS (Int.CL.7)
			G10L

Date d'achèvement de la recherche	Examinateur
29 août 2000	Van Doremalen, J

CATÉGORIE DES DOCUMENTS CITÉS

X : particulièrement pertinent à lui seul
Y : particulièrement pertinent en combinaison avec un autre document de la même catégorie
A : arrière-plan technologique
O : divulgation non-écrite
P : document intercalaire

T : théorie ou principe à la base de l'invention
E : document de brevet bénéficiant d'une date antérieure à la date de dépôt et qui n'a été publié qu'à cette date de dépôt ou qu'à une date postérieure.
D : cité dans la demande
L : cité pour d'autres raisons

& : membre de la même famille, document correspondant

EPO FORM 1503 12.99 (P04C14)