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

(11) EP 0 594 480 B1

(12) **FASCICULE DE BREVET EUROPEEN**

(45) Date de publication et mention de la délivrance du brevet:
18.08.1999 Bulletin 1999/33

(51) Int Cl.[6]: **G10L 3/00**

(21) Numéro de dépôt: **93402522.2**

(22) Date de dépôt: **13.10.1993**

(54) **Procédé de détection de la parole**

Verfahren zur Erkennung von Sprachsignalen

Speech detection method

(84) Etats contractants désignés:
DE GB

(30) Priorité: **21.10.1992 FR 9212582**

(43) Date de publication de la demande:
27.04.1994 Bulletin 1994/17

(73) Titulaire: **SEXTANT AVIONIQUE**
92360 Meudon-la-Forêt (FR)

(72) Inventeur: **Pastor, Dominique**
F-92402 Courbevoie Cedex (FR)

(74) Mandataire: **Chaverneff, Vladimir et al**
Thomson-CSF Propriété Intellectuelle,
13, Avenue du Président Salvador Allende
94117 Arcueil Cédex (FR)

(56) Documents cités:
EP-A- 0 335 521
EP-A- 0 459 363
US-A- 4 959 865

- **ICASSP 91 vol. 1 , 14 Mai 1991 , TORONTO pages 733 - 736 E.S. DERMATAS ET AL. 'Fast endpoint detection algorithm for isolated word recognition in office environment'**
- **ICASSP 85 vol. 4 , 26 Mars 1985 , TAMPA FLORIDA page 1838 M. CHUNG ET AL. 'Word boundary detection and speech recognition of noisy speech by means of iterative noise cancellation techniques'**
- **IEEE TRANS. ON ASSP vol. 27, no. 2 , Avril 1979 pages 113 - 120 S.F. BOLL 'Suppression of acoustic noise in speech using spectral subtraction'**

EP 0 594 480 B1

Printed by Jouve, 75001 PARIS (FR)

Description

[0001] La présente invention se rapporte à un procédé de détection de la parole.

[0002] Lorsqu'on cherche à déterminer le début et la fin effectifs de la parole, diverses solutions sont envisageables :

(1) On peut travailler sur l'amplitude instantanée par référence à un seuil déterminé expérimentalement et confirmer la détection de parole par une détection de voisement (voir article "La discrimination parole-bruit et ses applications" de V. PETIT/F. DUMONT, paru dans la Revue Technique THOMSON-CSF - Vol. 12 - N°4, déc. 1980).

(2) On peut aussi travailler sur l'énergie du signal total sur une tranche temporelle de durée T, en seuillant, toujours expérimentalement, cette énergie à l'aide d'histogrammes locaux, par exemple, et confirmer ensuite à l'aide d'une détection de voisement, ou du calcul de l'énergie minimale d'une voyelle. L'utilisation de l'énergie minimale d'une voyelle est une technique décrite dans le rapport "AMADEUS Version 1.0" de J.L. GAUVAIN du laboratoire LIMSI du CNRS.

(3) Les systèmes précédents permettent la détection de voisement, mais non pas du début et de la fin effectifs de la parole, c'est-à-dire la détection des sons fricatifs non voisés (/F/, /S/, /CH/) et des sons plosifs non voisés (P/, /T/, /Q/). Il faut donc les compléter par un algorithme de détection de ces fricatives. Une première technique peut consister en l'utilisation d'histogrammes locaux, comme le préconise l'article "Problème de détection des frontières de mots en présence de bruit additifs" de P. WACRENIER, paru dans le mémoire de D.E.A. de l'université de PARIS-SUD, Centre d'Orsay.

[0003] D'autres techniques voisines des précédentes et relativement proches de celle exposée ici, ont été présentées dans l'article "A Study of Endpoint Detection Algorithms in Adverse Conditions: Incidence on a DTW and HMM Recognizer" de J.C. JUNQUA/B. REAVES/B. MAK, lors du Congrès EUROSPEECH 1991.

[0004] Dans toutes ces approches, une grande part est faite à l'heuristique, et peu d'outils théoriques puissants sont utilisés.

[0005] Les travaux sur le débruitage de la parole similaires à ceux présentés ici sont beaucoup plus nombreux, et l'on citera en particulier le livre "Speech Enhancement" de J.S. LIM aux Editions Prentice-Hall Signal Processing Series "Suppression of Acoustic Noise in Speech Using Spectral Substraction" de S.F. BOLL, paru dans la revue IEEE Transactions on Acoustics, speech, and signal processing, vol. ASSP-27, N°2, Avril 1989, et "Noise Reduction For Speech Enhancement In Cars : Non-Linear Spectral Subtraction/Kalman Filtering" de P. LOCKWOOD, C. BAILLARGEAT, J. M. GILLOT, J. BOUDY, G. FAUCON paru dans la revue EUROSPEECH 91. On ne citera que des techniques de débruitage dans le domaine spectral, et il sera question par la suite de débruitage "spectral" par abus de langage.

[0006] Dans tous ces travaux, la relation étroite entre détection et débruitage n'est jamais réellement mise en évidence, sauf dans l'article "Suppression of Acoustic Noise in Speech Using Spectral Subtraction" précité, qui propose une solution empirique à ce problème.

[0007] Or, il est évident qu'un débruitage de la parole, lorsqu'on ne dispose pas de deux canaux d'enregistrements, nécessite l'utilisation de trames de bruit "pur", non polluées par la parole, ce qui nécessite de définir un outil de détection capable de distinguer entre bruit et bruit + parole.

[0008] L'état de la technique le plus proche est représenté par Dermatas et. al., "Fast Endpoint Detection Algorithm for Isolated Word Recognition in Office Environment", ICASSP '91, Vol. 1, pp. 733-736.

[0009] La présente invention a pour objet un procédé de détection et de débruitage de la parole qui permette de détecter le plus sûrement possible les débuts et fins effectifs de signaux de parole quels que soient les types de sons de parole, et qui permette de débruiter le plus efficacement possible les signaux ainsi détectés, même lorsque les caractéristiques statistiques du bruit affectant ces signaux varient fortement.

[0010] Le procédé de l'invention consiste à effectuer en milieu peu bruité une détection de trames voisées, et à détecter un noyau vocalique auquel on attache un intervalle de confiance.

[0011] En milieu bruité, après avoir effectué la détection d'au moins une trame voisée, on recherche des trames de bruit précédant cette trame voisée, on construit un modèle autorégressif de bruit et un spectre moyen de bruit, on blanchit par filtre réjecteur et on débruite par débruiteur spectral les trames précédant le voisement, on recherche le début effectif de la parole dans ces trames blanchies, on extrait des trames débruitées comprises entre le début effectif de la parole et la première trame voisée les vecteurs acoustiques utilisés par le système de reconnaissance vocale, tant que des trames voisées sont détectées, celles-ci sont débruitées puis paramétrisées en vue de leur reconnaissance (c'est-à-dire que l'on extrait les vecteurs acoustiques adaptés à la reconnaissance de ces trames), lorsqu'on ne détecte plus de trames voisées, on recherche la fin effective de la parole, on débruite puis on paramétrise les trames comprises entre la dernière trame voisée et la fin effective de la parole.

[0012] Dans toute la suite, lorsqu'il sera question de paramétrisation des trames, il faudra entendre que l'on extrait de la trame le vecteur acoustique (ou de manière équivalente, les paramètres acoustiques) utilisés par l'algorithme de reconnaissance.

[0013] Un exemple de tels paramètres acoustiques sont les coefficients cepstraux bien connus des spécialistes du traitement de la parole.

[0014] Dans toute la suite, on entendra par blanchiment, l'application du filtrage réjecteur calculé à partir du modèle autorégressif du bruit, et par débruitage, l'application du débruiteur spectral.

[0015] Blanchiment et débruitage spectral ne s'appliquent pas de manière séquentielle, mais en parallèle, le blanchiment permettant la détection de sons non voisés, le débruitage une amélioration de la qualité du signal vocal à reconnaître.

[0016] Ainsi, le procédé de l'invention est caractérisé par l'utilisation d'outils théoriques permettant une approche rigoureuse des problèmes de détection (voisement et fricatives), par sa grande adaptativité, car ce procédé est avant tout un procédé local au mot. Les caractéristiques statistiques du bruit peuvent évoluer dans le temps, le procédé restera capable de s'y adapter, par construction. Il est également caractérisé par l'élaboration d'expertises de détection à partir des résultats d'algorithmes de traitement du signal (on minimise ainsi le nombre de fausses alarmes, dues à la détection, en prenant en compte la nature particulière du signal de parole), par des processus de débruitage couplés à la détection de parole, par une approche en "temps réel", et ce, à tous les niveaux de l'analyse, par sa synergie avec d'autres techniques de traitement du signal vocal, par l'utilisation de deux débruiteurs différents :

* Filtrage de réjection, principalement utilisé pour la détection de fricatives, en vertu de ses propriétés de blanchiment.
* Filtrage de Wiener en particulier, utilisé pour débruiter le signal de parole en vue de sa reconnaissance. On peut aussi utiliser une soustraction spectrale.

[0017] Il faut donc distinguer dans le procédé de l'invention, trois niveaux de traitement :

- Le niveau "élémentaire" qui met en oeuvre des algorithmes de traitement du signal qui sont en fait les éléments de base de tous les traitements de niveau supérieur.

[0018] Ainsi, le niveau "élémentaire" de détection de voisement est un algorithme de calcul et de seuillage de la fonction de corrélation. Le résultat est expertisé par le niveau supérieur.

[0019] Ces traitements s'implantent sur processeurs de traitement du signal, par exemple du type DSP 96000.

- Le niveau intermédiaire d'expertise élabore des détections de voisements et de début de parole "intelligentes", compte tenu de la détection "brute" fournie par le niveau élémentaire. L'expertise peut faire appel à un langage informatique approprié, type Prolog.
- Le niveau "supérieur" ou utilisateur gère en temps réel les différents algorithmes de détection, de débruitage et d'analyse du signal vocal. Le langage C, par exemple, est approprié à l'implémentation de cette gestion.

[0020] L'invention est décrite en détail ci-dessous selon le plan suivant. On décrit d'abord l'algorithme qui permet d'enchaîner de façon appropriée les différentes techniques de traitement du signal et d'expertises nécessaires.

[0021] On supposera à ce niveau de traitement le plus élevé dans la hiérarchie de conception, que l'on dispose de méthodes fiables de détection et de débruitage, comportant tous les algorithmes de traitement de signal, toutes les expertises, nécessaires et suffisants. Cette description est donc très générale. Elle est même indépendante des algorithmes d'expertises et de traitement du signal décrits ci-après. Elle peut donc s'appliquer à d'autres techniques que celles décrites ici.

[0022] On décrit ensuite les expertises de détection de voisement, de début et fin de parole, à l'aide d'algorithmes de niveau élémentaire dont certains exemples sont cités.

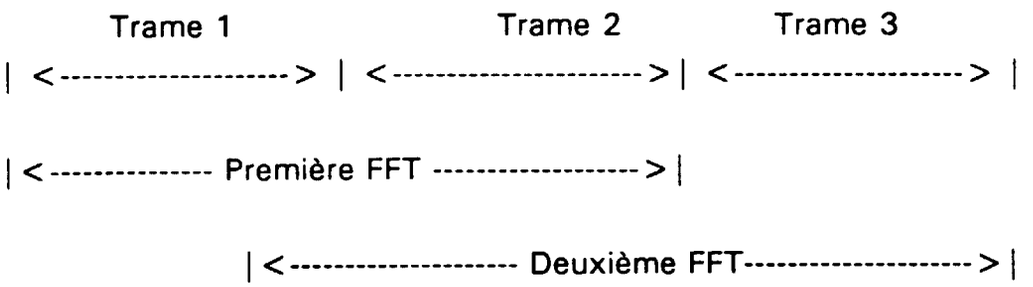
[0023] On décrit enfin les méthodes utilisées pour la détection et le débruitage de la parole.

[0024] Ce sont les résultats de ces techniques (Parole voisée, non voisée, ...) qui sont utilisés par les niveaux supérieurs de traitement

Conventions et vocabulaire employé.

[0025] On appellera trame, l'unité temporelle élémentaire de traitement. La durée d'une trame est classiquement de 12.8 ms, mais peut, bien entendu, avoir des valeurs (réalisations en langage mathématique) différentes. Les traitements font appel à des transformées de Fourier discrètes des signaux traités. Ces transformées de Fourier sont appliquées à l'ensemble des échantillons obtenus sur deux trames consécutives, ce qui correspond à effectuer une transformée de Fourier sur 25.6 ms.

[0026] Lorsque deux transformées de Fourier sont consécutives dans le temps, ces transformées sont calculées, non pas sur quatre trames consécutives, mais sur trois trames consécutives avec un recouvrement d'une trame. Ceci est illustré par le schéma suivant :



[0027] On décrit d'abord ici le fonctionnement de l'algorithme au niveau de conception le plus proche de l'utilisateur.

[0028] Le mode de mise en oeuvre préféré de l'invention est décrit ci-dessous en référence à l'analyse des signaux issus d'environnements avioniques très bruités, ce qui permet de disposer d'une information de départ qui est l'alternat micro qu'utilisent les pilotes. Cette information indique une zone temporelle proche du signal à traiter.

[0029] Cependant, cet alternat peut être plus ou moins rapproché du début effectif de la parole, et on ne peut donc y accorder qu'un faible crédit pour toute détection précise. Il va donc falloir préciser le début effectif de parole à partir de cette première information.

[0030] Dans un premier temps, on recherche la première trame voisée située aux alentours de cet alternat. Cette première trame voisée, est recherchée tout d'abord parmi les N1 trames qui précèdent l'alternat (N1 = environ 30 trames). Si cette trame voisée n'est pas trouvée parmi ces N1 trames, alors on recherche le voisement sur les trames qui suivent l'alternat, au fur et à mesure qu'elles se présentent.

[0031] Dès que la première trame voisée est trouvée par ce procédé, on va initialiser les débruiteurs. Pour cela, il faut mettre en évidence des trames constituées uniquement de bruit. Ces trames de bruit sont recherchées parmi les N2 trames qui précèdent la première trame voisée (N2 = environ 40 trames). En effet, chacune de ces N2 trames est soit :

* constituée de bruit seul
* constituée de bruit + respiration
* constituée de bruit + fricative ou occlusive non voisée.

[0032] L'hypothèse faite est que l'énergie du bruit est en moyenne inférieure à celle du bruit + respiration, elle-même inférieure à celle du bruit + fricative.

[0033] Donc, si on considère parmi les N2 trames, celle qui présente l'énergie la plus faible, il est fort probable que cette trame n'est constituée que de bruit.

[0034] A partir de la connaissance de cette trame, on recherche toutes celles qui sont compatibles avec elle, et compatibles 2 à 2, au sens donné plus loin, au paragraphe "Compatibilités entre énergies".

[0035] Lorsque les trames de bruit ont été détectées, on construit les deux modèles de bruit qui vont servir par la suite :

* Modèle autorégressif du bruit permettant de construire le filtrage réjecteur qui blanchit le bruit.
* Spectre moyen de bruit pour débruitage spectral.

[0036] Ces modèles sont décrits ci-dessous.

[0037] Les modèles de bruit étant construits, on blanchit (par filtre réjecteur et on débruite (par débruiteur spectral) les N3 trames qui précèdent le voisement et parmi lesquelles on va chercher le début effectif de parole (N3 = environ 30). Il va de soi que N3 est inférieur à N2. Cette détection se fait par détection de fricative et est décrite ci-dessous.

[0038] Lorsque le début de parole est connu, on débruite toutes les trames comprises entre le début de parole et la première trame voisée, puis on paramétrise ces trames en vue de leur reconnaissance. Au fur et à mesure que ces trames sont débruitées et paramétrisées, elles sont envoyées au système de reconnaissance.

[0039] Puisque le début effectif de parole est connu, on peut continuer à traiter les trames qui suivent la première trame voisée.

[0040] Chaque trame acquise n'est plus blanchie mais seulement débruitée, puis paramétrisée pour sa reconnaissance. On effectue sur chaque trame un test de voisement.

[0041] Si cette trame est voisée, le vecteur acoustique est effectivement envoyé à l'algorithme de reconnaissance.

[0042] Si elle n'est pas voisée, on cherche si elle est en fait la dernière trame du noyau vocalique en cours.

[0043] Si ce n'est pas la dernière trame du noyau vocalique, on acquiert une nouvelle trame et on réitère le procédé, jusqu'au moment où l'on trouve la dernière trame voisée.

[0044] Lorsque la dernière trame voisée est détectée, on blanchit les N4 trames qui suivent cette dernière trame voisée (N4 = environ 30 trames), puis on recherche la fin effective de la parole parmi ces N4 trames blanchies. Le procédé associé à cette détection est décrit ci-dessous.

[0045] Lorsque la fin effective de parole est détectée, les trames comprises entre la fin de voisement et cette fin de parole, sont débruitées, puis paramétrisées et envoyées au système de reconnaissance vocale pure en vue de leur traitement.

[0046] Lorsque la dernière trame de parole a été débruitée, paramétrisée et envoyée au système de reconnaissance, on réinitialise tous les paramètres de traitement, en vue du traitement de l'élocution suivante.

[0047] Comme on peut le voir, ce procédé est local à l'élocution traitée (c'est-à-dire qu'il traite chaque phrase ou chaque ensemble de mots sans "trou" entre mots), et permet donc d'être très adaptatif à tous changements de statistiques du bruit, d'autant plus que l'on utilise des algorithmes adaptatifs pour la modélisation auto-régressive du bruit, et des modèles théoriques relativement sophistiqués pour la détection des trames de bruit et la détection des fricatives.

[0048] En l'absence d'alternat, le procédé est lancé dès qu'un voisement est détecté.

[0049] Une simplification notable du procédé décrit ci-dessus est possible lorsque les signaux traités sont peu bruités. L'utilisation des algorithmes de débruitage et de blanchiment peut alors se révéler inutile, voire néfaste, lorsque le niveau de bruit est négligeable (ambiance laboratoire). Ce phénomène est connu, notamment dans le cas du débruitage, où débruiter un signal très peu bruité peut induire une déformation de la parole préjudiciable à une bonne reconnaissance.

Les simplifications résident :

- dans la suppression du débruitage spectral pour reconnaissance en vue d'éviter toute déformation de la parole, ne compensant pas le gain en rapport signal sur bruit que l'on pourrait obtenir par débruitage, et préjudiciable alors à une bonne reconnaissance.
- dans l'éventuelle suppression du filtre de blanchiment (et donc du calcul du modèle autorégressif du bruit, ce qui implique aussi la suppression du module de confirmation de bruit). Cette suppression n'est pas forcément nécessaire en milieu peu bruité. Des essais préalables sont préférables pour en décider.

[0050] On va maintenant exposer en détail les procédures d'expertise de détection de voisement et détection de fricative.

[0051] Ces procédures d'expertises font appel à des outils bien connus de traitement du signal et de détection, qui sont autant d'automates de base, dont l'aptitude est de décider de manière brute si la trame traitée est voisée ou non, est une trame de fricative non voisée ou de plosive non voisée...

[0052] L'expertise consiste à combiner les différents résultats obtenus à l'aide desdits outils, de manière à mettre en évidence des ensembles cohérents, formant le noyau vocalique par exemple, ou des blocs de sons fricatifs (plosifs), non voisés.

[0053] Par nature, le langage d'implémentation de telles procédures est de préférence le PROLOG.

[0054] A la différence du processus décrit ci-dessus, cette expertise est la même que le milieu soit bruité ou non.

[0055] Pour l'expertise de détection de voisement, on fait appel à un processus connu de détection de voisement, qui, pour une trame donnée, décide si cette trame est voisée ou non, en renvoyant la valeur du "pitch" associé à cette trame. Le "pitch" est la fréquence de répétition du motif de voisement. Cette valeur de pitch est nulle, s'il n'y a pas de voisement, et non nulle sinon.

[0056] Cette détection élémentaire de voisement se fait sans utiliser des résultats portant sur les trames précédentes, et sans présager du résultat portant sur les trames futures.

[0057] Comme un noyau vocalique peut être constitué de plusieurs segments voisés, séparés de trous non voisés, une expertise est nécessaire, afin de valider ou non un voisement.

[0058] On va maintenant exposer les règles générales de l'expertise.

Règle 1 : Entre deux trames voisées consécutives ou séparées d'un nombre relativement faible de trames (de l'ordre de trois ou quatre trames), les valeurs de pitch obtenues ne peuvent différer de plus d'un certain delta (environ ± 20 Hz en fonction du locuteur). Par contre, lorsque l'écart entre deux trames voisées excède un certain nombre de trames, la valeur de pitch peut évoluer très vite.

Règle 2 : Un noyau vocalique est constitué de trames voisées entrecoupées de trous. Ces trous doivent vérifier la condition suivante: la taille d'un trou ne doit pas excéder une taille maximale, qui peut être fonction du locuteur et surtout du vocabulaire (environ 40 trames). La taille du noyau est la somme du nombre de trames voisées et de la taille des trous de ce noyau.

Règle 3 : le début effectif du noyau vocalique est donné dès que la taille du noyau est suffisamment grande (environ 4 trames).

Règle 4 : la fin du noyau vocalique est déterminée par la dernière trame voisée suivie d'un trou excédant la taille

maximale permise pour un trou dans le noyau vocalique.

Déroulement de l'expertise

[0059] On utilise les règles précédentes de la manière exposée ci-dessous, et lorsqu'une valeur de pitch a été calculée.

Première partie de l'expertise :

[0060] On valide ou non la valeur du pitch calculée, en fonction de la valeur du pitch de la trame précédente et de la dernière valeur non nulle du pitch, et ce en fonction du nombre de trames séparant la trame actuellement traitée et celle du dernier pitch non nul. Ceci correspond à l'application de la règle 1.

Deuxième partie de l'expertise :

[0061] Cette deuxième partie de l'expertise se décompose suivant différents cas.

Cas 1 : Première trame voisée :

- On incrémente la taille possible du noyau, qui vaut donc 1
- Le début possible du noyau vocalique est donc la trame actuelle
- La fin possible du noyau vocalique est donc la trame actuelle

Cas 2 : La trame actuelle est voisée ainsi que la précédente.
On traite donc un segment voisé.

- On incrémente le nombre possible de trames voisées du noyau
- On incrémente la taille possible du noyau
- La fin possible du noyau peut être la trame actuelle qui est aussi la fin possible du segment.

Si la taille du noyau est suffisamment grande (environ quatre trames, comme précisé ci-dessus).
Et Si le début effectif du noyau vocalique n'est pas connu.
Alors :

- le début du noyau est la première trame détectée comme voisée.
  Ceci correspond à la mise en oeuvre de la règle 3.

Cas 3 : la trame actuelle n'est pas voisée, alors que la trame précédente l'est.
On est en train de traiter la première trame d'un trou.

- On incrémente la taille du trou, qui passe à 1

Cas 4 : La trame actuelle n'est pas voisée et la trame précédente non plus.
On est en train de traiter un trou.

- On incrémente la taille du trou.
  Si la taille du trou excède la taille maximale autorisée pour un trou du noyau vocalique,
  Alors :

  Si le début effectif de voisement est connu,
  Alors :
  la fin du noyau vocalique est la dernière trame voisée déterminée avant ce trou. On arrête l'expertise et on réinitialise toutes les données pour le traitement de la prochaine élocution. (cf règle 4)
  Si le début effectif de parole n'est toujours pas connu,
  Alors :
  On continue l'expertise sur les trames suivantes après réinitialisation de tous les paramètres utilisés, car ceux qui ont été actualisés précédemment ne sont pas valides.
  Sinon, ce trou fait peut-être partie du noyau vocalique et on ne peut pas encore prendre de décision définitive.

Cas 5 : La trame actuelle est voisée et la précédente ne l'est pas.

On vient de terminer un trou, et on commence un nouveau segment voisé.

- On incrémente le nombre de trames voisées du noyau.
- On incrémente la taille du noyau.

Si le trou que l'on vient de finir peut faire partie du noyau vocalique, (c'est-à-dire si sa taille est inférieure à la taille maximale autorisée pour un trou du noyau selon la règle 2).
Alors :

- On ajoute à la taille actuelle du noyau la taille de ce trou.
- On réinitialise la taille du trou, pour traitement des prochaines trames non voisées.

Si le début effectif du voisement n'est pas encore connu,
Et Si la taille du noyau est désormais suffisante(Règle 3),
Alors :

- le début du voisement est le début du segment voisé précédant le trou que l'on vient de terminer.

Sinon, ce trou ne peut pas faire partie du noyau vocalique :
Si le début effectif du voisement est connu,
Alors :

- la fin du noyau vocalique est la dernière trame voisée déterminée avant ce trou. On arrête l'expertise et on réinitialise toutes les données pour le traitement de la prochaine élocution. (cf règle 4).

Si le début effectif de voisement n'est toujours pas connu,
Alors :

- On continue l'expertise sur les trames suivantes après réinitialisation de tous les paramètres utilisés, car ceux qui ont été actualisés précédemment ne sont pas valides.

[0062] Cette procédure est utilisée à chaque trame, et après calcul du pitch associé à cette trame.
[0063] Expertise de détection de la parole non voisée.
[0064] On utilise ici un processus connu en soi de détection de parole non voisée.
[0065] Cette détection élémentaire de voisement se fait sans utiliser des résultats portant sur les trames précédentes, et sans présager du résultat portant sur les trames futures.
[0066] Des signaux de parole non voisés placés en début ou en fin d'élocution, peuvent être constitués :

- d'un seul segment fricatif comme dans "chaff"
- d'un segment fricatif suivi d'un segment occlusif comme dans "stop"
- d'un seul segment occlusif comme dans "parole"

[0067] Il y a donc possibilité de trous dans l'ensemble de trames non voisées.
[0068] De plus, de tels blocs fricatifs ne doivent pas être trop grands. Aussi, une expertise intervenant après la détection de ces sons est-elle nécessaire.
[0069] Dans la suite, par abus de langage, le terme fricatif se rapportera tout aussi bien à des fricatives non voisées qu'à des plosives non voisées.
[0070] Règles générales de l'expertise.
[0071] L'expertise exposée ici est similaire à celle décrite ci-dessus dans le cas du voisement. Les différences tiennent essentiellement dans la prise en compte des paramètres nouveaux que sont la distance entre le noyau vocalique et le bloc fricatif, et la taille du bloc fricatif.

Règle 1 : la distance entre le noyau vocalique et la première trame fricative détectée ne doit pas être trop grande (environ 15 trames maximum) .
Règle 2 : la taille d'un bloc fricatif ne doit pas être trop grande. Ceci signifie de manière équivalente, que la distance entre le noyau vocalique et la dernière trame détectée comme fricative ne doit pas être trop grande (environ 10

trames maximum).
Règle 3 la taille d'un trou dans un bloc fricatif ne doit pas excéder une taille maximale (environ 15 trames maximum). La taille totale du noyau est la somme du nombre de trames voisées et de la taille des trous dans ce noyau.
Règle 4 : le début effectif du bloc fricatif est déterminé dès que la taille d'un segment est devenue suffisante, et que la distance entre le noyau vocalique et la première trame de ce segment fricatif traité n'est pas trop grande, conformément à la règle 1. Le début effectif du bloc fricatif correspond à la première trame de ce segment.
Règle 5 : la fin du bloc fricatif est déterminée par la dernière trame du bloc fricatif suivie d'un trou excédant la taille maximale autorisée pour un trou dans le noyau vocalique, et lorsque la taille du bloc fricatif ainsi déterminé n'est pas trop grande conformément à la règle 2.

Déroulement de l'expertise.

[0072] Cette expertise est utilisée pour détecter les blocs fricatifs précédant le noyau vocalique ou le suivant. Le repère choisi dans cette expertise est donc le noyau vocalique.

[0073] Dans le cas de la détection d'un bloc fricatif précédant le noyau vocalique, le traitement se fait en partant de la première trame de voisement, donc en "remontant" dans le temps. Aussi, lorsque l'on dit qu'une trame i suit une trame j (précédemment traitée), il faut entendre par là : vis-à-vis de cette première trame du noyau vocalique. Dans la réalité, la trame j est chronologiquement postérieure à la trame i. Ce que l'on dénomme début du bloc fricatif dans l'expertise décrite ci-après, est en fait, chronologiquement, la fin de ce bloc, et ce que l'on appelle fin du bloc fricatif, est en fait le début chronologique de ce bloc. La distance entre noyau vocalique et trame détectée comme fricative est la distance entre la première trame du bloc voisé et cette trame de fricative.

[0074] Dans le cas de la détection d'un bloc fricatif situé après le noyau vocalique, le traitement se fait après la dernière trame voisée, et suit donc l'ordre chronologique naturel, et les termes de l'expertise sont parfaitement adéquats.

Cas 1 : Tant qu'il n'y a pas de détection de fricative, on est dans un trou qui suit le noyau vocalique et précède le bloc fricatif.

- On incrémente la distance entre le segment voisé et le bloc fricatif. Cette distance ainsi calculée est un minorant de la distance entre le bloc fricatif et le noyau vocalique. Cette distance sera figée dès que la première trame de fricative sera détectée.

Cas 2 : Première détection de fricative, On commence à traiter un segment fricatif.

- On initialise la taille du bloc fricatif à 1.
- On fige la distance entre le bloc voisé et le bloc fricatif.

Si la distance entre le noyau vocalique et le bloc fricatif n'est pas trop grande (conformément à la règle 2).
Alors :

- Le début possible du bloc fricatif peut être la trame actuelle.
- La fin possible du bloc fricatif peut être la trame actuelle.

Si la taille du bloc fricatif est suffisamment grande
Et Si le début effectif du bloc fricatif n'est pas encore connu,
alors :

- le début du noyau peut être confirmé.

On notera que ce **Si** (dans "**Si** la taille du bloc fricatif est suffisamment grande") est inutile si la taille minimale pour un bloc fricatif est supérieure à une trame, mais lorsqu'on cherche à détecter des occlusives en milieu bruité, celles-ci peuvent n'apparaître que sur la durée d'une seule trame. Il faut donc prendre alors la taille minimale d'un bloc fricatif égale à 1, et conserver cette condition.
Si la distance entre le noyau vocalique et le bloc fricatif est trop grande (cf règle 2).
Il n' y pas de bloc fricatif acceptable.

- On réinitialise pour le traitement de la prochaine élocution.
- On sort du traitement.

Comme le test sur la distance entre noyau vocalique et bloc fricatif est réalisé dès la première détection de fricative, il ne sera pas renouvelé dans les cas suivants, d'autant plus que si cette distance est ici trop grande, la procédure est arrêtée pour cette élocution.

Cas 3 : La trame actuelle et la précédente sont toutes les deux des trames de fricatives.
On est en train de traiter une trame qui se situe en plein dans un segment fricatif acceptable (situé à une distance correcte du noyau vocalique conformément à la règle 1).

- La fin possible du bloc fricatif est la trame actuelle.
- On incrémente la taille du bloc fricatif.

Si la taille du bloc fricatif est suffisamment grande (cf règle 4).
Et Si la taille de ce bloc n'est pas trop grande (cf règle 2).
Et Si le début effectif du bloc fricatif n'est pas encore connu,
alors :

- le début du noyau peut être confirmé comme étant le début de ce segment fricatif.

Cas 4: La trame actuelle n'est pas une fricative contrairement à la trame précédente.
On est en train de traiter la première trame d'un trou situé à l'intérieur du bloc fricatif.

- On incrémente la taille totale du trou (qui devient égale à 1).

Cas 5 : Ni la trame actuelle ni la précédente ne sont des trames de fricatives.
On est en train de traiter une trame située en plein dans un trou du bloc fricatif.

- On incrémente la taille totale du trou.

Si la taille actuelle du bloc fricatif augmentée de la taille du trou est supérieure à la taille maximale autorisée pour un bloc fricatif (règle 2).
Ou Si la taille du trou est trop grande.
Si le début du bloc fricatif est connu,
alors :

- La fin du bloc fricatif est la dernière trame détectée comme fricative.
- On réinitialise toutes les données de manière à traiter la prochaine élocution.

Sinon :

- on réinitialise toutes les données, même celles qui ont été précédemment actualisées, car elles ne sont plus valides. On traite alors la prochaine trame.

Sinon, ce trou fait peut-être partie du bloc fricatif et on ne peut pas encore prendre de décision définitive.
Cas 6 : La trame actuelle est une trame de fricative contrairement à la trame précédente.
On traite la première trame d'un segment fricatif situé après un trou.

- On incrémente la taille du bloc fricatif.

Si la taille actuelle du bloc fricatif augmentée de la taille du trou précédemment détecté est supérieure à la taille maximale autorisée pour un bloc fricatif,
Ou Si la taille du trou est trop grande,
alors :
Si le début du bloc fricatif est connu,
alors :

- La fin du bloc fricatif est alors la dernière trame détectée comme fricative.
- On réinitialise toutes les données de manière à traiter la prochaine élocution.

Sinon,

- On réinitialise toutes les données, même celles qui ont été précédemment actualisées, car elles ne sont pas valides. On traite alors la prochaine trame.

Sinon, (le trou fait partie du segment fricatif).

- La taille du bloc fricatif est augmentée de la taille du trou
- La taille du trou est réinitialisée à O
  Si la taille du bloc fricatif est suffisamment grande
  Et Si cette taille n'est pas trop grande
  Et Si le début effectif du bloc fricatif n'est pas connu
  Alors :
- Le début du noyau peut être confirmé.

Simplification dans le cas d'un milieu peu bruité.

[0075] Dans le cas où l'utilisateur estime que le milieu est insuffisamment bruité pour nécessiter les traitements sophistiqués précédents, il est possible, non seulement de simplifier l'expertise présentée ci-dessus, mais même de l'éliminer. Dans ce cas, la détection de parole se réduira à une simple détection du noyau vocalique auquel on attache un intervalle de confiance exprimé en nombre de trames, ce qui se révèle suffisant pour améliorer les performances d'un algorithme de reconnaissance vocale. Il est ainsi possible de débuter la reconnaissance une dizaine, voire une quinzaine de trames avant le début du noyau vocalique, et de l'achever, une dizaine, voire une quinzaine de trames après le noyau vocalique.

Algorithmes de Traitement du Signal.

[0076] Les procédures et méthodes de calcul décrites ci-après sont les constituants utilisés par les algorithmes d'expertises et de gestion. De telles fonctions sont avantageusement implantées sur un processeur de signaux et le langage utilisé est de préférence l'Assembleur.

[0077] Pour la détection de voisement en milieu peu bruité, une solution intéressante est le seuillage de l'A.M.D.F. (Average Magnitude Difference Function) dont la description peut être trouvée par exemple dans l'ouvrage "Traitement de la parole" de R. BOITE/M. KUNT paru aux éditions Presses Polytechniques Romandes.

[0078] L'AMDF est la fonction $D(k) = \Sigma_n |x(n+k) - x(n)|$. Cette fonction est bornée par la fonction de corrélation, selon : $D(k) \leq 2(\Gamma_x(0) - \Gamma_x(k))^{1/2}$. Cette fonction présente donc des "pics" vers le bas, et doit donc être seuillée comme la fonction de corrélation.

[0079] D'autres méthodes basées sur le calcul du spectre du signal sont envisageables, pour des résultats tout aussi acceptables (article "traitement de la parole" précité). Toutefois, il est intéressant d'utiliser la fonction AMDF, pour de simples questions de coûts de calcul.

[0080] En milieu bruité, la fonction AMDF est une distance entre le signal et sa forme retardée. Cependant, cette distance est une distance qui n'admet pas de produit scalaire associé, et qui ne permet donc pas d'introduire la notion de projection orthogonale. Or, dans un milieu bruité, la projection orthogonale du bruit peut être nulle, si l'axe de projection est bien choisi. L'AMDF n'est donc pas une solution adéquate en milieu bruité.

[0081] Le procédé de l'invention est alors basé sur la corrélation, car la corrélation est un produit scalaire et effectue une projection orthogonale du signal sur sa forme retardée. Cette méthode est, par là-même, plus robuste au bruit que d'autres techniques, telles l'AMDF. En effet, supposons que le signal observé soit $x(n) = s(n) + b(n)$ où $b(n)$ est un bruit blanc indépendant du signal utile $s(n)$. La fonction de corrélation est par définition :

$$\Gamma_x(k) = E[x(n)x(n-k)], \text{ donc } \Gamma_x(k) = E[s(n)s(n-k)] + E[b(n)b(n-k)]$$

$$= \Gamma_s(k) + \Gamma_b(k)$$

Comme le bruit est blanc :

$$\Gamma_x(0) = \Gamma_s(0) + \Gamma_b(0) \text{ et } \Gamma_x(k) = \Gamma_s(k) \text{ pour } k \neq 0 .$$

[0082] La blancheur du bruit en pratique n'est pas une hypothèse valide. Cependant, le résultat reste une bonne approximation dès que la fonction de corrélation du bruit décroît rapidement, et pour k suffisamment grand, comme dans le cas d'un bruit rose (bruit blanc filtré par un passe-bande), où la fonction de corrélation est un sinus cardinal,

donc pratiquement nulle dès que k est suffisamment grand.

[0083] On va décrire maintenant une procédure de calcul de pitch et de détection de pitch applicable aux milieux bruités comme aux milieux peu bruités.

[0084] Soit x(n) le signal traité où $n \in \{ 0, ..., N-1 \}$.

[0085] Dans le cas de l'AMDF, $r(k) = D(k) = \Sigma_n |x(n+k) - x(n)|$

[0086] Dans le cas de la corrélation, l'espérance mathématique permettant d'accéder à la fonction de corrélation ne peut qu'être estimée, de sorte que la fonction r(k) est : $r(k) = K \Sigma_{0 \leq n \leq N-1} x(n)x(n-k)$ où K est une constante de calibration.

[0087] Dans les deux cas, on obtient théoriquement la valeur du pitch en procédant comme suit : r(k) est maximale en k = 0. Si le second maximum de r(k) est obtenu en $k = k_0$, alors la valeur du voisement est $F_0 = F_e/k_0$ où F_e est la fréquence d'échantillonnage.

[0088] Cependant, cette description théorique doit être révisée en pratique.

[0089] En effet, si le signal n'est connu que sur les échantillons 0 à N-1, alors x(n-k) est pris nul tant que n n'est pas supérieur à k. Il n'y aura donc pas le même nombre de points de calcul d'une valeur k à l'autre. Par exemple, si la fourchette du pitch est prise égale à [100 Hz, 333 Hz], et ce, pour une fréquence d'échantillonnage de 10 KHz, l'indice k_1 correspondant à 100 Hz vaut :

$k_1 = F_e/F_0 = 10000/100 = 100$ et celui correspondant à 333 Hz vaut $k_2 = F_e/F_0 = 10000/333 = 30$.

[0090] Le calcul du pitch pour cette fourchette se fera donc de k = 30 à k = 100.

[0091] Si on dispose par exemple de 256 échantillons (2 trames de 12,8 ms échantillonnées à 10 KHz), le calcul de r(30) se fait de n = 30 à n = 128, soit sur 99 points et celui de r(100) de n = 100 à 128, soit sur 29 points.

[0092] Les calculs ne sont donc pas homogènes entre eux et n'ont pas la même validité.

[0093] Pour que le calcul soit correct, il faut que la fenêtre d'observation soit toujours la même quel que soit k. De sorte que si n-k est inférieure à 0, il faut avoir conservé en mémoire les valeurs passées du signal x(n), de manière à calculer la fonction r(k)sur autant de points, quel que soit k. La valeur de la constante K n'importe plus.

[0094] Ceci n'est préjudiciable au calcul du pitch que sur la première trame réellement voisée, puisque, dans ce cas, les échantillons utilisés pour le calcul sont issus d'une trame non voisée, et ne sont donc pas représentatifs du signal à traiter. Cependant, dès la troisième trame voisée consécutive, lorsqu'on travaille, par exemple, par trames de 128 points échantillonnés à 10 KHz, le calcul du pitch sera valide. Ceci suppose, de manière générale, qu'un voisement dure au minimum 3x12,8 ms, ce qui est une hypothèse réaliste. Cette hypothèse devra être prise en compte lors de l'expertise, et la durée minimale pour valider un segment voisé sera de 3x12,8 ms dans cette même expertise.

[0095] Cette fonction r(k) étant calculée, il s'agit ensuite de la seuiller. Le seuil est choisi expérimentalement, selon la dynamique des signaux traités. Ainsi, dans un exemple d'application, où la quantification se fait sur 16 bits, où la dynamique des échantillons n'excède pas $\pm$ 10000, et où les calculs se font pour N = 128 (Fréquence d'échantillonnage de 10 KHz), on a choisi Seuil = 750000. Mais rappelons que ces valeurs ne sont données qu'à titre d'exemple pour des applications particulières, et doivent être modifiées pour d'autres applications. En tout cas, cela ne change rien à la méthodologie décrite ci-dessus.

On va maintenant exposer le procédé de détection des trames de bruit.

[0096] En-dehors du noyau vocalique, les trames de signal que l'on peut rencontrer, sont de trois types :

1) bruit seul
2) bruit + fricative non voisée
3) bruit + respiration.

[0097] L'algorithme de détection vise à détecter le début et la fin de parole à partir d'une version blanchie du signal, tandis que l'algorithme de débruitage nécessite la connaissance du spectre moyen de bruit. Pour construire les modèles de bruit qui vont permettre de blanchir le signal de parole en vue de la détection des sons non voisés comme décrit ci-dessous, et pour débruiter le signal de parole, il est évident qu'il faut détecter les trames de bruit, et les confirmer en tant que telles. Cette recherche des trames de bruit se fait parmi un nombre de trames N_1 défini par l'utilisateur une fois pour toutes pour son application (par exemple pour $N_1 = 40$), ces N_1 trames étant situées avant le noyau vocalique.

[0098] Rappelons que cet algorithme permet la mise en oeuvre de modèles de bruit, et n'est donc pas utilisé lorsque l'utilisateur juge le niveau de bruit insuffisant.

[0099] On va d'abord définir les variables aléatoires gaussiennes "positives":

Une variable aléatoire X sera dite positive lorsque $Pr\{ X < 0 \} << 1$.

Soit X_0 la variable centrée normalisée associée à X. On a :

$Pr \{ X < 0 \} = Pr \{ X_0 < -m/\sigma \}$ où $m = E[X]$ et $\sigma^2 = E[(X-m)^2]$.

[0100] Dès que m/σ est suffisamment grand, X peut être considérée comme positive.

[0101] Lorsque X est gaussienne, on désigne par F(x) la fonction de répartition de la loi normale, et on a :

$Pr \{ X < 0 \} = F(-m/\sigma)$ pour $X \in N(m,\sigma^2)$

[0102] Une propriété essentielle immédiate est que la somme X de N variables gaussiennes positives indépendantes $X_i \in N(m_i ; \sigma_i^2)$ reste une variable gaussienne positive :

$$X = \Sigma_{1 \le i \le N} X_i \in N(\Sigma_{1 \le i \le N} m_i ; \Sigma_{1 \le i \le N} \sigma_i^2)$$

Résultat fondamental :

[0103] Si $X = X_1/X_2$ où X_1 et X_2 sont toutes deux des variables aléatoires gaussiennes, indépendantes, telles que $X_1 \in N(m_1 ; \sigma_1^2)$ et $X_2 \in N(m_2 ; \sigma_2^2)$, on pose $m = m_1/m_2$, $\alpha_1 = m_1/\sigma_1$, $\alpha_2 = m_2/\sigma_2$.

[0104] Lorsque α_1 et α_2 sont suffisamment grands pour pouvoir supposer X_1 et X_2 positives, la densité de probabilité $f_X(x)$ de $X = X_1/X_2$ peut alors être approchée par :

$$f_X(x) = (2\pi)^{-1/2} \alpha_1 \alpha_2 m \frac{\alpha_1^2 x + \alpha_2^2 m}{\left(\alpha_1^2 x^2 + \alpha_2^2 m^2\right)^{3/2}} e^{-\frac{\alpha_1^2 \alpha_2^2 (x-m)^2}{2\left(\alpha_1^2 x^2 + \alpha_2^2 m^2\right)}} . U(x)$$

où U(x) est la fonction indicatrice de $\mathbf{R}^+$:
$U(x) = 1$ si $x \le 0$ et $U(x) = 0$ si $x < 0$
Dans toute la suite, on posera :

$$f(x, y) \Big| \alpha, \beta) = (2\pi)^{-1/2} \alpha\beta\, y \frac{\alpha^2 x + \beta^2 y}{\left(\alpha^2 x^2 + \beta^2 y^2\right)^{3/2}} \cdot e^{-\frac{\alpha^2 \beta^2 (x-y)^2}{2\left(\alpha^2 x^2 + \beta^2 y^2\right)}}$$

de sorte que : $f_X(x) = f(x,m|\alpha_1,\alpha_2).U(x)$ Soit

$$h(x,y|\alpha,\beta) = \alpha\beta \frac{x - y}{\left(\alpha^2 x^2 + \beta^2 y^2\right)^{1/2}}$$

On pose $P(x,y|\alpha,\beta) = F[h(x,y|\alpha,\beta)]$. On a alors :
Pr $\{X < x\} = P(x,m|\alpha_1,\alpha_2)$ $f(x,y|\alpha,\beta) = \partial P(x,y|\alpha,\beta)/\partial x$ et $f(x,y|\alpha_1,\alpha_2) = \partial P(x,m|\alpha_1,\alpha_2)/\partial x$
Cas particulier : $\alpha = \beta$. On posera: $f_\alpha(x,y) = f(x,y|\alpha,\beta)$, $h_\alpha(x,y) = h(x,y|\alpha,\beta)$ et $P_\alpha(x,y) = P(x,y|\alpha,\beta)$

[0105] On va décrire ci-dessous quelques modèles de base de variables gaussiennes "positives" utilisables dans la suite.

(1) Signal à énergie déterministe : Soient les échantillons x(0),...x(N-1) d'un signal quelconque, dont l'énergie est déterministe et constante, ou approximée par une énergie déterministe ou constante.
On a donc $U = \Sigma_{0 \le n \le N-1} x(n)^2 \in N(N\mu, 0)$ où

$$\mu = (1/N)\Sigma_{0 \le n \le N-1} x(n)^2$$

Prenons comme exemple le signal $x(n) = A\cos(n+\theta)$ où θ est équiréparti entre $[0,2\pi]$. Pour N suffisamment grand, on a :

(1/N) $\Sigma_{0\leq n\leq N-1}x(n)^2$ # $E[x(n)^2] = A^2/2$. Pour N assez grand, U peut être assimilé à $NA^2/2$ et donc à une énergie constante.

(2) Processus Blanc Gaussien : Soit un processus blanc et gaussien x(n) tel que $\sigma_x^2 = E[x(n)^2]$.

Pour N suffisamment grand,

$$U = \Sigma_{0\leq n\leq N-1}x(n)^2 \in N(N\sigma_x^2 ; 2N\sigma_x^4).$$

Le paramètre α est $\alpha = (N/2)^{1/2}$

(3) Processus Gaussien Bande Etroite : Le bruit x(n) est issu de l'échantillonnage du processus x(t), lui-même issu du filtrage d'un bruit blanc gaussien b(t) par un filtre passe bande h(t) : x(t) = (h∗b)(t), en supposant que la fonction de transfert du filtre h(t) est :

$$H(f) = U_{[-(f0-B/2,-f0+B/2]}(f) + U_{[f0-B/2,f0+B/2]}(f),$$

où U désigne la fonction caractéristique de l'intervalle en indice et f_0 la fréquence centrale du filtre.

On a donc $U \in N(N\sigma_x^2, 2\sigma_x^4\Sigma_{0\leq i\leq N-1,0\leq j\leq N-1}g_{f0,B,Te}(i-j)^2)$ avec $g_{f0,B,Te}(k) = \cos(2\pi kf_0T_e)\sin_c(\pi kBT_e)$

Le paramètre α et $\alpha = N/[2\Sigma_{0\leq i\leq N-1,0\leq j\leq N-1}g_{f0,B,Te}(i-j)^2]^{1/2}$

(4) Sous-échantillonnage d'un processus gaussien : Ce modèle est plus pratique que théorique. Si la fonction de corrélation est inconnue, on sait cependant que : $\lim_{k\to+\infty}\Gamma_x(k) = 0$. Donc, pour k assez grand tel que $k > k_0$, la fonction de corrélation tend vers 0. Aussi, au lieu de traiter la suite d'échantillons x(0)...x(N-1), peut-on traiter la sous-suite $x(0), x(k_0), x(2k_0),...$, et l'énergie associée à cette suite reste une variable aléatoire positive gaussienne, à condition qu'il reste dans cette sous-suite suffisamment de points pour pouvoir appliquer les approximations dues au théorème central-limite.

Compatibilité entre énergies.

Soient $C_1 = N(m_1,\sigma_1^2)$ et $C_2 = N(m_2,\sigma_2^2)$

On pose : $m = m_1/m_2$, $\alpha_1 = m_1/\sigma_1$ et $\alpha_2 = m_2/\sigma_2$.

α_1 et α_2 sont suffisamment grands pour que les variables aléatoires de C_1 et C_2 puissent être considérées comme des variables aléatoires positives.

Soit (U,V) où (U,V) appartient à $(C_1UC_2)X(C_1UC_2)$.

Comme précédemment, U et V sont supposées indépendantes.

On pose $U \equiv V \Leftrightarrow (U,V) \in (C_1XC_1)U(C_2UC_2)$.

Soit (u,v) une valeur du couple (U,V). Si x = u/v, x est une valeur de la variable aléatoire X = U/V.

Soit s 1.

$1/s < x < s \Leftrightarrow$ On décide que $U \equiv V$ est vrai, ce qui sera la décision $D = D_1$

$x < 1/s$ ou $x > s \Leftrightarrow$ On décide que $U \equiv V$ est faux, ce qui sera la décision $D = D_2$ Cette règle de décision est donc associée à 2 hypothèses :

$H_1 \Leftrightarrow U \equiv V$ est vrai , $H_2 \Leftrightarrow U \equiv V$ est faux .

On posera I = [1/s,s].

La règle de détection s'exprime encore selon : $x \in I \Leftrightarrow D = D_1$,

$x \in R - I \Leftrightarrow D = D_2$

[0106] On dira que u et v sont compatibles lorsque la décision $D = D_1$ sera prise.

[0107] Cette règle de décision admet une probabilité de décision correcte, dont l'expression dépendra en fait de la valeur des probabilités $Pr\{H_1\}$ et $Pr\{H_2\}$.

[0108] Or, ces probabilités ne sont en général pas connues en pratique.

[0109] On préfère alors une approche du type Neyman-Pearson, puisque la règle de décision se réduit à deux hypothèses, en cherchant à assurer une certaine valeur fixée a priori pour la probabilité de fausse alarme qui est :

$$P_{fa} = Pr\{ D_1 \mid H_2 \} = P(s,m|\alpha_1,\alpha_2) - P(1/s,m|\alpha_1,\alpha_2)$$

[0110] Le choix des modèles des signaux et des bruits détermine α_1 et α_2. Nous allons voir qu'alors m apparaît comme homogène à un rapport signal sur bruit qui sera fixé de manière heuristique. Le seuil est alors fixé de manière à assurer une certaine valeur de P_{fa}.

Cas particulier : $\alpha_1 = \alpha_2 = \alpha$. Il vient alors : $P_{fa} = P_\alpha(s,m) - P_\alpha(1/s,m)$

Compatibilité d'un ensemble de valeurs :
Soit { u_1, ... , u_n } un ensemble de valeurs de variables aléatoires gaussiennes positives. On dira que ces valeurs sont compatibles entre elles, si et seulement si les u_i sont compatibles 2 à 2.
Modèles du signal et du bruit utilisés par le procédé de l'invention.

[0111] Afin d'appliquer les procédures correspondant aux rappels théoriques précédents, il faut fixer un modèle du bruit et du signal. Nous utiliserons l'exemple suivant. Ce modèle est régi par les hypothèses suivantes :

Hypothèse 1 : Nous supposons ne pas connaître le signal utile dans sa forme, mais nous ferons l'hypothèse suivante : pour les valeurs s(0),...,s(N-1) de s(n), l'énergie $S = (1/N)\Sigma_{0 \le n \le N-1}\, s(n)^2$ est bornée par μ_s^2, et ce, dès que N est suffisamment grand, de sorte que :

$$S = \Sigma_{0 \le n \le N-1}\, s(n)^2 > N\mu_s^2$$

Hypothèse 2 : Le signal utile est perturbé par un bruit additif noté x(n), que l'on suppose gaussien et en bande étroite. On suppose que le processus x(n) traité est obtenu par filtrage bande étroite d'un bruit blanc gaussien. La fonction de corrélation d'un tel processus est alors :

$$\Gamma_x(k) = \Gamma_x(0)\cos(2\pi k f_0 T_e)\mathrm{sinc}(\pi k B T_e).$$

Si on considère N échantillons x(n) de ce bruit, et qu'on pose :
$g_{f0,B,Te}(k) = \cos(2\pi k f_0 T_e)\sin_c(\pi k B T_e)$, on a :

$$V = (1/N)\Sigma_{0 \le n \le N-1}\, x(n)^2 \in N(N\sigma_x^2,$$

$$2\sigma_x^4 \Sigma_{0 \le i \le N\,1, 0 \le j \le N-1} g_{f0,B,Te}(i-j)^2)$$

Le paramètre α de cette variable est α

$$= N/[2\,\Sigma_{0 \le i \le N-1, 0 \le j \le N-1} g_{f0,B,Te}^{(i-j)^2}]^{1/2}$$

Hypothèse 3 : Les signaux s(n) et x(n) sont alors supposés indépendants. On suppose que l'indépendance entre s(n) et x(n) implique la décorrélation au sens temporel du terme, c'est-à-dire que l'on peut écrire :

$$c = \frac{\Sigma_{0 \le n \le N-1} s(n)x(n)}{(\Sigma_{0 \le n \le N-1} s(n)^2)^{1/2}\,(\Sigma_{0 \le n \le N-1}\, x(n)^2)^{1/2}} = 0$$

Ce coefficient de corrélation n'est que l'expression dans le domaine temporel du coefficient de corrélation spatial défini par :
$E[s(n)x(n)]/(E[s(n)^2]E[x(n)^2])^{1/2}$ lorsque les processus sont ergodiques.
Soit u(n) = s(n) + x(n) le signal total, et $U = \Sigma_{0 \le n \le N-1}\, u(n)^2$.
On peut alors approximer U par : $U = \Sigma_{0 \le n \le N-1}\, s(n)^2 + \Sigma_{0 \le n \le N-1} x(n)^2$
Comme on a : $\Sigma_{0 \le n \le N-1}\, s(n)^2 \ge \mu_s^2$,
on aura : $U \ge N\mu_s^2 + \Sigma_{0 \le n \le N-1}\, x(n)^2$.
Hypothèse 4 : Comme nous supposons que le signal présente une énergie moyenne bornée, nous supposerons qu'un algorithme capable de détecter une énergie μ_s^2, sera capable de détecter tout signal d'énergie supérieure.
Compte tenu des hypothèses précédentes, on définit la classe C_1 comme étant la classe des énergies lorsque le signal utile est présent. Selon l'hypothèse 3, $U \ge N\mu_s^2 + \Sigma_{0 \le n \le N-1}\, x(n)^2$, et selon l'hypothèse 4, si on détecte l'énergie $N\mu_s^2 + \Sigma_{0 \le n \le N-1}\, x(n)^2$, on saura détecter aussi l'énergie totale U.
D'après l'hypothèse 2,
$N\mu_s^2 + \Sigma_{0 \le n \le N-1}\, x(n)^2 \in N(N\mu_s^2 + N\sigma_x^2,$
$2\sigma_x^4 \Sigma_{0 \le i \le N-1, 0 \le j \le N-1} g_{f0,B,Te}^{(i-j)^2})$.
Donc $C_1 = N(N\mu_s^2 + N\sigma_x^2,\ 2\sigma_x^4 \Sigma_{0 \le i \le N-1, 0 \le j \le N-1} g_{f0,B,Te}^{(i-j)^2})$

et le paramètre α de cette variable vaut
$\alpha_1 = N(1 + r)/[2\ \Sigma_{0 \le i \le N-1, 0 \le j \le N-1} g_{f0,B,Te}(i-j)^2]^{1/2}$,
où $r = \mu_s^2/\sigma_x^2$ représente le rapport signal à bruit.
C_2 est la classe des énergies correspondant au bruit seul. D'après l'hypothèse 2, si les échantillons de bruit sont $x(0),\ldots,x(M-1)$,
il vient $V = (1/M)\Sigma_{0 \le n \le M-1}\ x(n)^2 \in N(M\sigma_x^2,$
$2\sigma_x^4\ \Sigma_{0 \le i \le M-1, 0 \le j \le M-1} g_{f0,B,Te}(i-j)^2)$.
Le paramètre α de cette variable est :
$\alpha_2 = M/[2\Sigma_{0 \le i \le M-1, 0 \le j \le M-1} g_{f0,B,Te}(i-j)^2]^{1/2}$
On a donc : $C_1 = N(m_1,\sigma_1^2)$ et $C_2 = N(m_2,\sigma_2^2)$,
avec : $m_1 = N\mu_s^2 + N\sigma_x^2$, $m_2 = M\sigma_x^2$,
$\sigma_1 = \sigma_x^2[2\Sigma_{0 \le i \le N-1, 0 \le j \le N-1} g_{f0,B,Te}(i-j)^2]^{1/2}$ et
$\sigma_2 = \sigma_x^2[2\Sigma_{0 \le i \le M-1, 0 \le j \le M-1} g_{f0,B,Te}(i-j)^2]^{1/2}$.
D'où $m = m_1/m_2 = (N/M)(1+r)$,
$\alpha_1 = m_1/\sigma_1 = N(1-r)/[2\Sigma_{0 \le i \le N-1, 0 \le j \le N-1} g_{f0,B,Te}(i-j)^2]^{1/2}$ et
$\sigma_2 = m_2/\sigma_2 = M/[2\Sigma_{0 \le i \le M-1, 0 \le j \le M-1} g_{f0,B,Te}(i-j)^2]^{1/2}$.

[0112] On remarquera que :

- si le bruit d'origine est blanc et gaussien, les hypothèses précédentes restent encore valables. Il suffit de remarquer qu'alors $g_{f0,B,Te}(k) = \delta_0(k)$. Les formules précédentes s'en trouvent simplifiées :
  $C_1 = N(m_1,\sigma_1^2)$ et $C_2 = N(m_2,\sigma_2^2)$,
  avec : $m_1 = N\mu_s^2 + N\sigma_x^2$, $m_2 = M\sigma_x^2$, $\sigma_1^2 = 2N\sigma_x^4$ et $\sigma_2 = 2M\sigma_x^4$.
  D'où $m = m_1/m_2 = (N/M)(1+r)$,
  $\alpha_1 = m_1/\sigma_1 = (1+r)(N/2)^{1/2}$ et
  $\alpha_2 = m_2/\sigma_2 = (M/2)^{1/2}$.
  Il est possible de tendre vers un tel modèle en sous-échantillonnant le bruit, et ne prenant du bruit qu'un échantillon sur k_0 échantillons où k_0 est tel que: : $\forall\ k > k_0$, $\Gamma_x(k) \to 0$.
- la notion de compatibilité entre énergies ne se met en place que conditionnellement à la connaissance a priori du paramètre m, donc du rapport signal à bruit r. Celui-ci peut être fixé de manière heuristique à partir de mesures préliminaires des rapports signaux à bruit que présentent les signaux que l'on ne veut pas détecter par l'algorithme de confirmation de bruit, ou fixé de manière péremptoire. La seconde solution est utilisée de préférence. En effet, l'objet de ce traitement vise à mettre en évidence, non pas toutes les trames de bruit, mais seulement quelques unes présentant une forte probabilité de n'être constituées que de bruit. On a donc tout intérêt à ce que l'algorithme soit très sélectif. Cette sélectivité s'obtient en jouant sur la valeur de la probabilité de fausse alarme que l'on décide d'assurer et qui sera donc choisie très faible (la sélectivité maximale étant établie pour $P_{FA} = 0$, ce qui conduit à un seuil nul et à aucune détection de bruit, ce qui est le cas extrême et absurde). Mais cette sélectivité s'obtient aussi par le choix de r : choisi trop grand, on risque de considérer des énergies comme représentatives du bruit, alors que ce sont des énergies de respiration, par exemple, présentant un rapport signal à bruit inférieur à r. A contrario, choisir un r trop petit peut limiter la P_{FA} accessible, qui serait alors trop forte pour être acceptable.

[0113] Compte tenu des modèles précédents, et le calcul du seuil ayant été fait, on applique alors l'algorithme suivant de détection et de confirmation de bruit, basé essentiellement sur la notion de compatibilité, telle que décrite ci-dessus.

[0114] La recherche et la confirmation des trames de bruit se fait parmi un nombre de trames N_1 défini par l'utilisateur une fois pour toute pour son application (par exemple $N_1 = 40$), ces trames étant situées avant le noyau vocalique. On fait l'hypothèse suivante: l'énergie des trames de bruit seul est en moyenne inférieure à celle des trames de bruit+respiration et de bruit signal. La trame présentant l'énergie minimale parmi les N_1 trames, est donc supposée n'être constituée que de bruit. On cherche alors toutes les trames compatibles avec cette trame au sens rappelé ci-dessus, en utilisant les modèles précités.

[0115] L'algorithme de détection de bruit va chercher, parmi un ensemble de trames $T_1,\ldots,T_n$, celles qui peuvent être considérées comme du bruit.

[0116] Soient $E(T_1),\ldots,E(T_n)$, les énergies de ces trames, calculées sous la forme : $E(T_i)= \Sigma_{0 \le n \le N-1}\ u(n)^2$ où $u(n)$ sont les N échantillons constituant la trame T_i.

[0117] On fait l'hypothèse suivante : la trame présentant l'énergie la plus faible est une trame de bruit. Soit T_{i0} cette trame.

[0118] L'algorithme se déroule comme suit :

L'ensemble des trames de bruit est initialisé: Bruit=$\{T_{i0}\}$
Pour i décrivant $\{ E(T_1),\ldots,E(T_n)\} - \{E(T_{i0})\}$

Faire
Si $E(T_i)$ est compatible avec chaque élément de Bruit:
Bruit = Bruit U $\{E(T_i)\}$
Fin pour

Modèle Autorégressif du bruit.

[0119] Puisque l'algorithme de confirmation de bruit fournit un certain nombre de trames qui peuvent être considérées comme du bruit avec une très forte probabilité, on cherche à construire, à partir de la donnée des échantillons temporels, un modèle autorégressif du bruit.

[0120] Si x(n) désigne les échantillons de bruit, on modélise x(n) sous la forme : $x(n) = \Sigma_{1 \leq i \leq p} a_i\, x(n-i) + b(n)$, où p est l'ordre du modèle, les a_i, les coefficients du modèle à déterminer et b(n) le bruit de modélisation, supposé blanc et gaussien si on suit une approche par maximum de vraisemblance.

[0121] Ce type de modélisation est largement décrit dans la littérature notamment dans "Spectrum Analysis - A Modern Perspective", de S.M. KAY et S.L. MARPLE Jr, paru dans Proceedings of the IEEE, Vol. 69, N° 11, novembre 1981..

[0122] Quant aux algorithmes de calcul du modèle, de nombreuses méthodes sont disponibles (Burg, Levinson-Durbin, Kalman, Fast Kalman...).

[0123] On utilise de préférence les méthodes du type Kalman et Fast Kalman, voir articles "Le Filtrage Adaptatif Transverse" de O.MACCHI/ M.BELLANGER paru dans la revue Traitement du Signal, Vol. 5, N° 3, 1988 et "Analyse des signaux et filtrage numérique adaptatif" de M.BELLANGER paru dans la Collection CNET-ENST, MASSON, qui présentent de très bonnes performances temps réel. Mais ce choix n'est pas le seul possible. L'ordre du filtre est par exemple choisi égal à 12, sans que cette valeur soit limitative.

Filtrage réjecteur

[0124] Soit u(n) = s(n) + x(n) le signal total, composé du signal de parole s(n) et du bruit x(n) .
Soit le filtre $H(z) = 1 - \Sigma_{1 \leq i \leq p} a_i z^{-i}$.
Appliqué au signal U(z),on obtient H(z)U(z) = H(z)S(z) + H(z)X(z).
Or : H(z)X(z) = B(z) = > H(z)U(z) = H(z)S(z) + B(z)

[0125] Le filtre réjecteur H(z) blanchit le signal, de sorte que le signal en sortie de ce filtre est un signal de parole (filtré donc déformé), additionné d'un bruit généralement blanc et gaussien.

[0126] Le signal obtenu est en fait impropre à la reconnaissance, car le filtre réjecteur déforme le signal de parole originel.

[0127] Cependant, le signal obtenu étant perturbé par un bruit pratiquement blanc et gaussien, il s'ensuit que ce signal est très intéressant pour effectuer une détection du signal s(n) selon la théorie exposée ci-dessous, selon laquelle on garde le signal large bande obtenu, ou on le filtre préalablement dans la bande des fricatives, comme décrit ci-dessous (cf "détection de fricatives").

[0128] C'est pour cette raison que l'on utilise ce filtrage réjecteur après modélisation auto-régressive du bruit.
Spectre moyen de bruit.

[0129] Comme l'on dispose d'un certain nombre de trames confirmées comme étant des trames de bruit, on peut alors calculer un spectre moyen de ce bruit, de manière à implanter un filtrage spectral, du type soustraction spectrale ou filtrage de WIENER.

[0130] On choisit par exemple le filtrage de WIENER. Aussi, a-t-on besoin de calculer $C_{XX}(f) = E[|X(f)|^2]$ qui représente le spectre moyen de bruit. Comme les calculs sont numériques, on n'a accès qu'à des FFT de signaux numériques pondérés par une fenêtre de pondération. De plus, la moyenne spatiale ne peut qu'être approximée.

[0131] Soient $X_1(n)...,X_M(n)$ les FFT des M trames de bruit confirmées comme telles, ces FFT étant obtenues par pondération du signal temporel initial par une fenêtre d'apodisation adéquate.

$C_{XX}(f) = E[|X(f)|^2]$ est approximé par :

$$\hat{C}_{XX}(n) = M_{XX}(n) = (1/M)\, \Sigma_{1 \leq i \leq M+1}\, |X_i(n)|^2$$

[0132] Les performances de cet estimateur sont données par exemple dans le livre "Digital signal processing" de L. RABINER/C.M.RADER paru chez IEEE Press.

[0133] Pour ce qui est du filtre de Wiener, on rappelle ci-dessous quelques résultats classiques, explicités notamment dans l'ouvrage "Speech Enhancement" de J.S. LIM paru aux Editions Prentice-Hall Signal Processing Series.

[0134] Soit u(t) = s(t) + x(t) le signal total observé, où s(t) désigne le signal utile (de parole) et x(t) le bruit.
Dans le domaine fréquentiel, on obtient : U(f) = S(f) + X(f), avec des notations évidentes.

[0135] On cherche alors le filtre H(f), de sorte que le signal $\hat{S}(f) = H(f)U(f)$ soit le plus proche de S(f) au sens de la

norme L_2. On cherche donc H(f) minimisant : $E[|S(f)-\hat{S}(f)|^2]$.
On démontre alors que : $H(f) = 1 - (C_{XX}(f)/C_{UU}(f))$ où
$C_{XX}(f) = E[|X(f)|^2]$ et $C_{UU}(f) = E[|U(f)^2|]$.

[0136] Ce type de filtre, parce que son expression est directement fréquentielle, est particulièrement intéressant à appliquer dès que la paramétrisation est basée sur le calcul du spectre.
Implémentation par corrélogramme lissé.

[0137] En pratique, C_{XX} et C_{UU} ne sont pas accessibles. On ne peut que les estimer. Une procédure d'estimation de $C_{XX}(f)$ a été décrite ci-dessus.

[0138] C_{UU} est le spectre moyen du signal total u(n) dont l'on ne dispose que sur une seule et unique trame. De plus, cette trame doit être paramétrisée de manière à pouvoir intervenir dans le processus de reconnaissance. Il n'est donc pas question d'effectuer une moyenne quelconque du signal u(n) d'autant plus que le signal de parole est un signal particulièrement non-stationnaire.

[0139] Il faut donc construire, à partir de la donnée de u(n), une estimation de $C_{UU}(n)$. On utilise alors le corrélogramme lissé.

[0140] On estime alors $C_{UU}(n)$ par: $\hat{C}_{UU}(k) = \Sigma_{0 \le n \le N-1} F(k-n)|X(n)|^2$
où F est une fenêtre de lissage construite comme suit, et N le nombre de points permettant le calcul des FFT : N = 256 points par exemple.
On choisit une fenêtre de lissage dans le domaine temporel :
$f(n) = a_0 + a_1\cos(2\pi n/N) + a_2\cos(4\pi n/N)$. Ces fenêtres sont largement décrites dans l'article précité : "On the Use of Windows for Hamming Analysis with the Discrete Fourier Transform de F.J.HARRIS paru dans Proceedings of the IEEE, Vol.66, N° 1, January 1978.
La fonction F(k) est alors simplement la Transformée de Fourier Discrète de f(n).
$\hat{C}_{UU}(k) = \Sigma_{0 \le n \le N-1} F(k-n)|X(n)|^2$ apparaît comme une convolution discrète entre F(k) et $V(k) = |X(k)^2|$, de sorte que $\hat{C}_{UU} = F * V$
Soit $\hat{c}_{UU}$ la FFT^{-1} de $\hat{C}_{UU}$. $\hat{c}_{UU}(k) = f(k)v(k)$ où v(k) est la FFT^{-1} de V(k).
On calcule donc $\hat{C}_{UU}(k)$ selon l'algorithme dit de corrélogramme lissé suivant :

(1) Calcul de v(k) par FFT inverse de $V(n) = |X(n)|^2$
(2) Calcul du produit f.v
(3) FFT directe du produit f.v qui aboutit à $\hat{C}_{UU}$

[0141] Plutôt que d'appliquer le même estimateur pour le bruit et le signal total, le procédé de l'invention applique l'algorithme du corrélogramme lissé précédent au spectre moyen de bruit $M_{XX}(n)$.
$\hat{C}_{XX}(k)$ est donc obtenu par :

$$\hat{C}_{XX}(k) = \Sigma_{0 \le n \le N-1} F(k-n)|M_{XX}(n)|^2$$

Le filtre de Wiener H(f) est donc estimé par la suite des valeurs :
$\hat{H}(n) = 1 - (\hat{C}_{XX}(n)/\hat{C}_{UU}(n))$
Le signal débruité a pour spectre : $\hat{S}(n) = \hat{H}(n)U(n)$
Une FFT^{-1} peut permettre, éventuellement, de récupérer le signal temporel débruité.

[0142] Le spectre débruité $\hat{S}(n)$ obtenu est le spectre utilisé pour la paramétrisation en vue de la reconnaissance de la trame.

[0143] Pour effectuer la détection des signaux non voisés, on utilise également les procédures décrites ci-dessus, puisque l'on dispose d'énergies représentatives du bruit (voir ci-dessus l'algorithme de détection du bruit).
Détection d'activité.
Soient $C_1 = N(m_1,\sigma_1^2)$ et $C_2 = N(m_2,\sigma_2^2)$.
Puisqu'on dispose d'un algorithme capable de mettre en évidence des valeurs de variables aléatoires appartenant à la même classe, de la classe C_2 (par exemple), et ce, avec une très faible probabilité d'erreur, il devient alors beaucoup plus facile de décider, par observation du couple U/V, si U appartient à la classe C_1 ou à la classe C_2.
Il y a donc deux hypothèses distinctes possibles,
$H_1 \Leftrightarrow U \in C_1$ et $H_2 \Leftrightarrow U \in C_2$
correspondant à deux décisions possibles distinctes :
$D = D_1 \Leftrightarrow$ Décision $U \in C_1$, notée "$U \in C_1$"
$D = D_2 \Leftrightarrow$ Décision $U \in C_2$, notée "$U \in C_2$"
Décision optimale.
On pose : $m = m_1/m_2$, $\alpha_1 = m_1/\sigma_1$ et $\alpha_2 = m_2/\sigma_2$.

Soit un couple (U,V) de variables aléatoires, où on suppose que $V \in C_2$ et $U \in C_1 U C_2$. U et V sont supposées indépendantes. En observant la variable X = U/V , on cherche à prendre une décision entre les deux suivantes possibles : "$C_1 X C_2$" , "$C_2 X C_2$".

On a donc deux hypothèses : $H_1 \Leftrightarrow U \in C_1$, $H_2 \Leftrightarrow U \in C_2$.

Soit $p = Pr\{ U \in C_1 \}$.

La règle de décision s'exprime sous la forme suivante :

$x > s \Leftrightarrow U \in C_1$, $x < s \Leftrightarrow U \in C_2$

La probabilité de décision correcte $P_c(s,ml\alpha_1,\alpha_2)$ est alors :

$P_c(s,ml\alpha_1,\alpha_2) = p[1-P(s,ml\alpha_1,\alpha_2)] + (1-p)P(s,1l\alpha_2,\alpha_2)$

où $p = Pr\{ U \in C_1\}$.

Le seuil optimal est celui pour lequel $P_c(s,ml\alpha_1,\alpha_2)$ est maximal. On résout donc l'équation :

$\partial P_c(s,ml\alpha_1,\alpha_2)/\partial s = 0 \Leftrightarrow pf(s,ml\alpha_1,\alpha_2) - (1-p)f(s,1l\alpha_2,\alpha_2) = 0$

Approche type Neyman-Pearson

Dans l'approche précédente, on supposait connaître la probabilité p. Lorsque cette probabilité est inconnue, on peut utiliser une approche type Neyman-Pearson.

On définit les probabilités de non détection et de fausse alarme :

$P_{nd} = \{ x < s \mid H_1 \}$ et $P_{fa} = \{ x > s \mid H_2 \}$

On a : $P_{nd} = P(s,1l\alpha_2,\alpha_2)$ et $P_{fa} = 1-P(s,ml\alpha_1,\alpha_2)$

On se fixe alors P_{fa} ou P_{nd}, pour déterminer la valeur du seuil.

[0144] Afin d'appliquer la détection d'activité telle que décrite ci-dessus au cas de la parole, il est nécessaire d'établir un modèle énergétique des signaux non voisés compatible avec les hypothèses qui président au bon fonctionnement des procédés décrits ci-dessus. On cherche donc un modèle des énergies des fricatives non voisées /F/, /S/, /CH/ et des plosives non voisées /P/, /T/, /Q/, qui permettent d'obtenir des énergies dont la loi statistique est approximativement une gaussienne.

Modèle 1.

[0145] Les sons /F/,/S/,/CH/ se situent spectralement dans une bande de fréquence qui s'étale d'environ 4 KHz à plus de 5KHz. Les sons /P/, /T/, /Q/ en tant que phénomènes courts dans le temps, s'étalent sur une bande plus large. Dans la bande choisie, on suppose que le spectre de ces sons fricatifs est relativement plat, de sorte que le signal fricatif dans cette bande peut se modéliser par un signal bande étroite. Ceci peut être réaliste dans certains cas pratiques sans avoir recours au blanchiment décrit ci-dessus. Cependant, dans la plupart des cas, il est judicieux de travailler sur un signal blanchi de manière à assurer un modèle de bruit à bande étroite convenable.

[0146] En acceptant un tel modèle de bruit à bande étroite, on a donc à traiter le rapport de deux énergies qui peut être traité par les procédés décrits ci-dessus.

[0147] Soient s(n) le signal de parole dans la bande étudiée et x(n) le bruit dans cette même bande. Les signaux s (n) et x(n) sont supposés indépendants.

[0148] La classe C_1 correspond à l'énergie du signal total u(n) = s(n) + x(n) observé sur N points, la classe C_2 correspond à l'énergie V du bruit seul observé sur M points.

[0149] Les signaux étant gaussiens et indépendants, u(n) est un signal lui-même gaussien, de sorte que :

$U = \Sigma_{0 \le n \le N-1} u(n)^2 \in N(N\sigma_u^2 , 2\sigma_u^4 \Sigma_{0 \le i \le N-1 , 0 \le j} \le N-1 g_{f0,B}(i-j)^2)$

De même :

$V = \Sigma_{0 \le n \le M-1} y(n)^2 \in N(M\sigma_x^2, 2\sigma_x^2 \Sigma_{0 \le i \le M-1, 0 \le j \le M-1} g_{f0,B}(1-j)^2)$, où y(n) désigne, on le rappelle, une autre valeur du bruit x(n) sur une tranche temporelle différente de celle où on observe u(n).

On peut donc appliquer les résultats théoriques ci-dessus avec :

$m = (N/M)\sigma_u 2/\sigma_x^2$,

$\alpha_1 = N/(2\Sigma_{0 \le i \le N-1 , 0 \le j \le N-1} g_{f0,B}(i-j)^2)^{1/2})$,

$\alpha_2 = M/(2\Sigma_{0 \le i \le M-1, 0 \le j \le M-1} g_{f0,B}(i-j)^2)^{1/2})$

[0150] On remarquera que $m = (N/M)(1+r)$ où $r = \sigma_s^2/\sigma_x^2$ désigne finalement le rapport signal sur bruit.

[0151] Pour achever complètement la résolution de ce problème, il faut pouvoir connaître le rapport signal sur bruit r ainsi que la probabilité de présence p du signal utile. Ce qui paraît être ici une limitation est commun aux deux autres modèles traités ci-dessous.

Modèle 2.

[0152] Comme dans le cas du modèle 1, on cherche à détecter uniquement les fricatives non voisées, donc à détecter un signal dans une bande particulière.

[0153] Ici, le modèle du signal fricatif n'est pas le même que précédemment. On suppose que les fricatives présentent

l'énergie minimale $\mu_s^2 = \Sigma_{0 \le n \le N-1}\, s(n)^2$ connue, grâce par exemple à un apprentissage, ou estimée.

[0154] Le son voisé est indépendant du bruit x(n) qui est ici gaussien bande étroite.

[0155] Si y(n), pour n compris entre 0 et M-1, désigne une autre valeur du bruit x(n) sur une tranche temporelle distincte de celle où est observé le signal total u(n) = s(n) + x(n), on aura :

$V = \Sigma_{0 \le n \le M-1}\, y(n)^2 \in N(M\sigma_x^2, 2Tr(C_{x,M}2))$ où $C_{x,M}$ désigne la matrice de corrélation du M-uplet : ${}^t(y(0),....y(M-1))$

[0156] En ce qui concerne l'énergie $U = \Sigma\, 0 \le n \le N\text{-}1\ u(n)^2$ du signal total, celle-ci peut s'exprimer selon :

$U = N\mu_s^2 + \Sigma_{0 \le n \le N-1} x(n)^2$

[0157] Ce résultat s'obtient en supposant que l'indépendance entre s(n) et x(n) s'exprime par la décorrélation au sens temporel du terme, c'est-à -dire que l'on peut écrire :

$$c = \frac{\Sigma\, 0 \le n \le N\text{-}1 s(n)x(n)}{(\Sigma_{0 \le n \le N-1} s(n)^2)^{1/2}\ (\Sigma_{0 \le n \le N-1} x(n)^2)^{1/2}} = 0$$

[0158] Comme $V' = \Sigma_{0 \le n \le N-1}\, x(n)^2 \in N(N\sigma_x^2, 2Tr(C_{x,N}^2))$ où $C_{x,N}$ désigne la matrice de corrélation du N-uplet : ${}^t(x(0),...,x(N-1))$, on a alors:

$U = \mu_s^2\ \mu_s^2\ {}_{0 \le n \le N-1}\, x(n)^2 \in N(N+ \Sigma + N\sigma_x^2 , 2Tr(C_{x,N}^2))$ On peut donc appliquer les résultats théoriques ci-dessus avec :

$C_1 = N(N\mu_s^2 + N\sigma_x^2 , 2Tr(C_{x,N}^2))$, $C_2 = N(M\sigma_x^2 , 2Tr(C_{x,M}^2))$ $m = (N/M)(1 + \mu_s^2/\sigma_x^2)$,

$\alpha_1 = N(\mu_s^2 + \sigma_x^2)/(2Tr(C_{x,N}^2))^{1/2}$, $\alpha_2 = M\sigma_x^2/(2Tr(C_{x,M}^2))^{1/2}$,

On remarquera que $m = (N/M)(1+r)$ où $r = \mu_s^2/\sigma_x^2$ désigne finalement le rapport signal sur bruit. La même remarque que celle du Modèle 1, concernant le rapport signal sur bruit r et la probabilité p de présence du signal utile, est valable ici.

Modèle 3.

[0159] Dans ce modèle, on cherche à effectuer une détection de tous les signaux non voisés, avec une hypothèse bruit blanc gaussien.

[0160] Le modèle signal bande étroite utilisé précédemment, n'est donc plus valable. On ne peut donc que supposer avoir affaire à un signal large bande dont on connaît l'énergie minimale $\mu\ s^2$.

Il vient donc :

$C_1 = N(N\mu_s^2 + N\sigma_x^2, 2N\sigma_x^4)$, $C_2 = N(M\sigma_x^2 , 2M\sigma_x^4)$

$m = (N/M)(1+r)$, avec $r = \mu_s^2/\sigma_x^2$

$\alpha_1 = (1+r)(N/2)^{1/2}$, $\alpha_2 = (M/2)^{1/2}$,

[0161] Pour utiliser ce modèle, le bruit doit être blanc gaussien. Si le bruit d'origine n'est pas blanc, on peut s'approcher de ce modèle en sous-échantillonnant en fait le signal observé, c'est-à-dire en ne considérant qu'un échantillon sur 2, 3, voire plus, suivant la fonction d'autocorrélation du bruit, et en supposant que le signal de parole ainsi sous-échantillonné présente encore une énergie décelable. Mais on peut aussi, et cela est préférable, utiliser cet algorithme sur un signal blanchi par filtre réjecteur, puisqu'alors le bruit résiduel est approximativement blanc et gaussien.

[0162] Les remarques précédentes concernant la valeur a priori du rapport signal à bruit et de la probabilité de présence du signal utile, restent encore et toujours valables.

Algorithmes de détection des sons non voisés.

[0163] En utilisant les modèles précédents, on expose ci-dessous deux algorithmes de détection des sons non voisés.

Algorithme 1 :

[0164] Disposant d'énergies représentatives de bruit, on peut moyenner ces énergies de sorte que l'on obtient une énergie de "référence" de bruit. Soir E_0 cette énergie. Pour N_3 trames T_1, ..., T_n qui précèdent la première trame voisée, on procède comme suit :

[0165] Soient $E(T_1)$, ..., $E(T_n)$, les énergies de ces trames, calculées sous la forme $E(T_i) = \Sigma_{0 \le n \le N-1} u(n)^2$ où u(n) sont les N échantillons constituant la trame T_i.

Pour $E(T_i)$ décrivant $\{E(T_1), ..., E(T_n)\}$

Faire

[0166] Si $E(T_i)$ est compatible avec E_0 (Décision sur la valeur de $E(T_i)/E_0$).

Détection sur la trame T_i.

Fin pour.

Algorithme 2 :

[0167] Cet algorithme est une variante du précédent. On utilise pour E_0, soit l'énergie moyenne des trames détectées comme du bruit, soit la valeur de l'énergie la plus faible de toutes les trames détectées comme du bruit.

[0168] Puis on procède comme suit :

Pour $E(T_i)$ décrivant $\{E(T_1), ..., E(T_n)\}$.

Faire

[0169] Si $E(T_i)$ est compatible avec E_0 (Décision sur la valeur de $E(T_i)/E_0$).

[0170] Détection sur la trame T_i.

Sinon $E_0 = E(T_i)$.

Fin pour

[0171] Le rapport signal à bruit r peut être estimé ou fixé de manière heuristique, à condition d'effectuer quelques mesures expérimentales préalables, caractéristiques du domaine d'application, de manière à fixer un ordre de grandeur du rapport signal sur bruit que présentent les fricatives dans la bande choisie.

[0172] La probabilité p de présence de la parole non voisée est, elle-aussi, une donnée heuristique, qui module la sélectivité de l'algorithme, au même titre d'ailleurs que le rapport signal à bruit. Cette donnée peut être estimée suivant le vocabulaire utilisé et le nombre de trames sur lequel se fait la recherche de sons non voisés.

Simplification dans le cas d'un milieu faiblement bruité.

[0173] Dans le cas d'un milieu faiblement bruité, pour lequel aucun modèle de bruit n'a été déterminé, en vertu des simplifications proposées ci-dessus, la théorie rappelée précédemment justifie l'utilisation d'un seuil, qui n'est pas lié de manière bijective au rapport signal à bruit, mais qui sera fixé de manière totalement empirique.

[0174] Une alternative intéressante pour des milieux où le bruit est négligeable, est de se contenter de la détection de voisement, d'éliminer la détection des sons non voisés, et de fixer le début de parole à quelques trames avant le noyau vocalique (environ 15 trames) et la fin de parole à quelques trames après la fin du noyau vocalique (environ 15 trames).

Revendications

1. Procédé de détection de la parole pour l'utilisation dans un système de reconnaissance vocale dans des signaux bruités, caractérisé par le fait qu'après avoir effectué dans ces signaux la détection d'au moins une trame voisée, on recherche des trames de bruit précédant cette trame voisée, on construit un modèle autorégressif de bruit et un spectre moyen de bruit, on blanchit par filtre réjecteur et on débruite par débruiteur spectral les trames précédant le voisement, on recherche le début effectif de la parole dans ces trames blanchies, on extrait des trames débruitées comprises entre le début effectif de la parole et la première trame voisée les vecteurs acoustiques utilisés par le système de reconnaissance vocale, tant que des trames voisées sont détectées, celles-ci sont débruitées puis paramétrisées en vue de leur reconnaissance, lorsqu'on ne détecte plus de trames voisées, on recherche la fin effective de la parole dans les trames blanchies qui suivent la dernière trame voisée, on débruite puis on paramétrise les trames comprises entre la dernière trame voisée et la fin effective de la parole.

2. Procédé selon la revendication 1, caractérisé par le fait que le blanchiment réalisé par filtrage réjecteur est calculé à partir du modèle autorégressif du bruit.

3. Procédé selon la revendication 2, caractérisé par le fait que lorsque la dernière trame de parole a été paramétrisée, on réinitialise tous les paramètres de traitement.

4. Procédé selon l'une des revendications précédentes, caractérisé par le fait que les trames de signaux à traiter sont traitées par transformées de Fourier, et que lorsque deux transformées sont consécutives dans le temps, elles sont calculées sur trois trames consécutives avec recouvrement d'une trame.

5. Procédé selon l'une des revendications précédentes, caractérisé par le fait que la détection de voisement se fait, pour chaque trame, à l'aide de la valeur du "pitch" associé à cette trame.

6. Procédé selon la revendication 5, caractérisé par le fait que l'on valide le calcul du pitch après avoir reconnu au

moint trois trames voisées, soit 3 x 12,8 ms.

7. Procédé selon la revendication 5 ou 6, caractérisé par le fait que le calcul du pitch est fait à partir de la corrélation du signal avec sa forme retardée.

8. Procédé selon l'une des revendications 5 à 7, caractérisé par le fait que la détection de sons non voisés se fait par seuillage.

9. Procédé selon l'une des revendications précédentes, caractérisé par le fait que pour détecter de la parole non voisée, on examine la distance entre le noyau vocalique et le bloc fricatif, et la taille de ce bloc fricatif.

10. Procédé selon l'une des revendications précédentes, caractérisé par le fait que le spectre moyen de bruit est obtenu par filtrage de Wiener.

11. Procédé selon la revendication 10, caractérisé par le fait que l'on applique l'algorithme du corrélogramme lissé au spectre moyen de bruit.

12. Procédé selon la revendication 1, caractérisé en outre par le fait qu'il est déterminé si le milieu est bruité ou peu bruité, et que dans le dernier cas on effectue uniquement une détection de trames voisées, et une détection de noyau vocalique auquel on attache un intervalle de confiance.

Patentansprüche

1. Verfahren zur Erfassung der Sprache in verrauschten Signalen zur Verwendung in einem Stimmerkennungssystem, dadurch gekennzeichnet, daß nach der Ausführung der Erfassung wenigstens eines stimmhaften Rahmens in diesen Signalen Rauschrahmen, die diesem stimmhaften Rahmen folgen, gesucht werden, ein autoregressives Rauschmodell und ein mittleres Rauschspektrum konstruiert werden, mittels eines Zurückweisungsfilters eine Reduzierung vorgenommen wird und die der Stimmhaftigkeit vorhergehenden Rahmen durch eine spektrale Rauschentfernungseinrichtung von Rauschen befreit werden, der effektive Anfang der Sprache in diesen reduzierten Rahmen gesucht wird, aus den vom Rauschen befreiten Rahmen, die zwischen dem effektiven Anfang der Sprache und dem ersten stimmhaften Rahmen enthalten sind, die akustischen Vektoren extrahiert werden, die vom Stimmerkennungssystem verwendet werden, solange die stimmhaften Rahmen erfaßt werden, wobei diese vom Rauschen befreit werden und dann im Hinblick auf ihre Erkennung parametrisiert werden, wenn keine stimmhaften Rahmen mehr erfaßt werden, das effektive Ende der Sprache in den reduzierten Rahmen, die dem letzten stimmhaften Rahmen folgen, gesucht wird, und die zwischen dem letzten stimmhaften Rahmen und dem effektiven Ende der Sprache enthaltenen Rahmen vom Rauschen befreit und dann parametrisiert werden.

2. Verfahren nach Anspruch 1, dadurch gekennzeichnet, daß die Reduzierung, die durch Zurückweisungsfilterung ausgeführt wird, anhand des autoregressiven Rauschmodells berechnet wird.

3. Verfahren nach Anspruch 2, dadurch gekennzeichnet, daß dann, wenn der letzte Sprachrahmen parametrisiert worden ist, sämtliche Verarbeitungsparameter neu initialisiert werden.

4. Verfahren nach einem der vorangehenden Ansprüche, dadurch gekennzeichnet, daß die zu verarbeitenden Signalrahmen durch Fourier-Transformationen verarbeitet werden, und, wenn zwei Transformationen zeitlich aufeinanderfolgen, diese für drei aufeinanderfolgende Rahmen mit Überlappung eines Rahmens berechnet werden.

5. Verfahren nach einem der vorangehenden Ansprüche, dadurch gekennzeichnet, daß die Erfassung der Stimmhaftigkeit für jeden Rahmen mit Hilfe des Tonhöhenwertes ("Pitch"-Wertes), der diesem Rahmen zugeordnet ist, erfolgt.

6. Verfahren nach Anspruch 5, dadurch gekennzeichnet, daß die Berechnung der Tonhöhe validiert wird, nachdem wenigstens drei stimmhafte Rahmen erkannt worden sind, also nach 3 × 12,8 ms.

7. Verfahren nach Anspruch 5 oder 6, dadurch gekennzeichnet, daß die Berechnung der Tonhöhe anhand der Korrelation des Signals mit seiner verzögerten Form erfolgt.

8. Verfahren nach einem der Ansprüche 5 bis 7, dadurch gekennzeichnet, daß die Erfassung von nicht stimmhaften Tönen durch Vergleichen mit Schwellenwerten erfolgt.

9. Verfahren nach einem der vorangehenden Ansprüche, dadurch gekennzeichnet, daß zur Erfassung von nicht stimmhafter Sprache der Abstand zwischen dem Vokalkern und dem Frikativblock und die Größe dieses Frikativblocks untersucht werden.

10. Verfahren nach einem der vorangehenden Ansprüche, dadurch gekennzeichnet, daß das mittlere Rauschspektrum durch Wiener-Filterung erhalten wird.

11. Verfahren nach Anspruch 10, dadurch gekennzeichnet, daß auf das mittlere Rauschspektrum der Algorithmus des gleichmäßigen Korrelogramms angewendet wird.

12. Verfahren nach Anspruch 1, außerdem dadurch gekennzeichnet, daß bestimmt wird, ob das Medium verrauscht oder wenig verrauscht ist, und daß im letzteren Fall ausschließlich eine Erfassung stimmhafter Rahmen und eine Erfassung des Vokalkerns, dem ein Vertrauensintervall hinzugefügt wird, ausgeführt werden.

Claims

1. Method of detecting speech, for use in a voice recognition system, in noisy signals, characterized in that, after having carried out, in these signals, the detection of at least one voiced frame, noise frames preceding this voiced frame are sought, an autoregressive model of noise and a mean noise spectrum are constructed, the frames preceding the voicing are bleached by rejector filter and noise is removed by spectral noise removal, the actual start of speech is sought in these bleached frames, from the noise-removed frames lying between the actual start of speech and the first voiced frame are extracted the acoustic vectors used by the voice recognition system, as long as voiced frames are detected, the latter have the noise removed then are parameterized for the purpose of recognizing them, when no more voiced frames are detected, the actual end of speech is sought in the bleached frames following the last voiced frame, the frames lying between the last voiced frame and the actual end of speech have the noise removed and are then parameterized.

2. Method according to Claim 1, characterized in that the bleaching carried out by a rejector filtering is calculated on the basis of the autoregressive model of the noise.

3. Method according to Claim 2, characterized in that, when the last speech frame has been parameterized, all the processing parameters are reinitialized.

4. Method according to one of the preceding claims, characterized in that the frames of signals to be processed are processed by Fourier transforms, and in that, when two transforms are consecutive in time, they are calculated over three consecutive frames with an overlap of one frame.

5. Method according to one of the preceding claims, characterized in that the detection of voicing is done, for each frame, with the aid of the value of the pitch associated with this frame.

6. Method according to Claim 5, characterized in that the calculation of the pitch is validated after having recognized at least three voiced frames, i.e. 3 x 12.8 ms.

7. Method according to Claim 5 or 6, characterized in that the calculation of the pitch is done from the correlation of the signal with its delayed form.

8. Method according to one of Claims 5 to 7, characterized in that the detection of unvoiced sounds is done by thresholding.

9. Method according to one of the preceding claims, characterized in that, in order to detect unvoiced speech, the distance between the vocal kernel and the fricative block, and the size of this fricative block, are examined.

10. Method according to one of the preceding claims, characterized in that the mean noise spectrum is obtained by Wiener filtering.

11. Method according to Claim 10, characterized in that the algorithm of the smooth correlogram is applied to the mean noise spectrum.

12. Method according to Claim 1, characterized in that, furthermore, it is determined whether the medium is noisy or slightly noisy, and in that, in the latter case, only a detection of voiced frames, and a vocal kernel detection to which a confidence interval is attached, are carried out.