(19) RÉPUBLIQUE FRANÇAISE

INSTITUT NATIONAL DE LA PROPRIÉTÉ INDUSTRIELLE

COURBEVOIE

(11) **N° de publication :** **3 026 542**

(à n'utiliser que pour les commandes de reproduction)

(21) **N° d'enregistrement national :** **14 59325**

(51) Int Cl[8] : ***G 10 L 15/22*** *(2016.01)*

(12) DEMANDE DE BREVET D'INVENTION A1

(22) **Date de dépôt :** 30.09.14.

(30) **Priorité :**

(43) **Date de mise à la disposition du public de la demande :** 01.04.16 Bulletin 16/13.

(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 :**

( ) **Demande(s) d'extension :**

(71) **Demandeur(s) :** *XBRAINSOFT Société par actions simplifiée* — FR.

(72) **Inventeur(s) :** RENARD GREGORY.

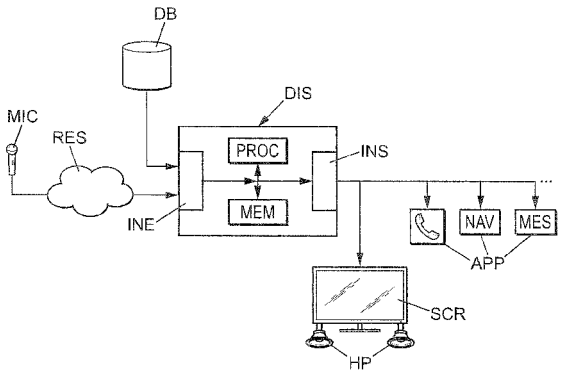
(73) **Titulaire(s) :** XBRAINSOFT Société par actions simplifiée.

(74) **Mandataire(s) :** CABINET PLASSERAUD.

(54) RECONNAISSANCE VOCALE PERFECTIONNEE.

(57) L'invention concerne une reconnaissance vocale, dans laquelle une unité de traitement (PROC, MEM) détermine un niveau d'intelligibilité d'un commande vocale prononcée par un utilisateur et, en cas d'inintelligibilité de la commande, transmet à une interface homme/machine (SCR, HP) un signal de demande de clarification par l'utilisateur d'une partie au moins de la commande vocale. En particulier, on prévoit les étapes dans lesquelles:

- l'unité de traitement évalue un score d'intelligibilité de la commande vocale, et
- l'unité de traitement transmet à l'interface homme/machine un signal de demande de clarification en fonction du score d'intelligibilité.



FR 3 026 542 - A1

Reconnaissance vocale perfectionnée

La présente invention concerne le domaine de la reconnaissance vocale.

L'invention trouve une application avantageuse mais non limitée à la conception d'assistants personnels intelligents, notamment dans le cadre de la commande par instructions vocales d'ordinateurs de bord de véhicules automobiles, par exemple pour des applications de navigation par GPS, ou encore pour des applications de lancement d'appels téléphoniques en mains libres ou de messageries, ou encore de consultation de documents.

Dans ce qui suit, les instructions vocales sont appelées « commandes vocales », le terme « instructions » étant réservé aux instructions de l'application informatique qui interprète les commandes vocales et qui pilote l'exécution des applications correspondantes (assister la navigation jusqu'à un point précis, lancer un appel téléphonique vers un contact précis, etc.).

On connait des assistants personnels intelligents qui comprennent les commandes vocales prononcées par des utilisateurs et qui répondent à leurs requêtes. Leur fonctionnement se base sur une reconnaissance vocale avancée, avec un traitement du langage oral naturel, en procédant à une synthèse vocale. Dans ce type de procédé, les mots prononcés par l'utilisateur sont comparés à une base et la succession des mots la plus probable par rapport à la comparaison de chaque mot est retenue comme étant la commande vocale de l'utilisateur.

Or souvent, cette commande vocale, une fois ainsi construite, doit être interprétée d'un bloc. Dans le cas où elle ne correspond à aucune instruction d'une application à lancer. L'assistant personnel indique simplement qu'il n'a pas compris la commande et demande de répéter. Les mots répétés ne conduisent pas à une amélioration de la compréhension par l'assistant, ce qui bloque définitivement l'interactivité et procure à l'utilisateur un mauvais ressenti de l'interface.

La présente invention vient améliorer la situation.

Elle propose en particulier d'améliorer le traitement de reconnaissance vocale pour qu'une interface homme/machine puisse guider l'utilisateur dans l'énoncé d'une commande vocale correctement interprétable par l'assistant personnel.

L'invention vise un procédé mis en œuvre par des moyens informatiques, de reconnaissance vocale, dans lequel une unité de traitement détermine un niveau d'intelligibilité d'un commande vocale prononcée par un utilisateur et, en cas d'inintelligibilité de la commande, transmet à une interface homme/machine un signal de demande de clarification par l'utilisateur d'une partie au moins de la commande vocale. Plus particulièrement, le procédé comporte alors les étapes dans lesquelles:

- l'unité de traitement évalue un score d'intelligibilité de la commande vocale, et
- l'unité de traitement transmet à l'interface homme/machine un signal de demande de clarification en fonction de ce score d'intelligibilité.

Ainsi, selon le score précité, si la commande vocale ne correspond pas à une instruction complète pouvant être exécutée, il est laissé à l'utilisateur la possibilité de clarifier ou compléter la commande vocale. L'utilisateur est alors guidé dans cette clarification en fonction du score qui a été attribué à la compréhension de ce que l'utilisateur a prononcé. L'invention va alors au-delà d'une simple compréhension de chaque mot de la commande. Ici, l'unité de traitement vérifie la cohérence de l'ensemble de la commande et sa complétude. Par exemple, si la commande prononcée n'est pas complète, l'unité de traitement guide l'utilisateur pour la rendre plus complète.

Ainsi, dans une forme de réalisation, l'unité de traitement exécute une application comportant un menu à rubriques arborescentes. Cette application peut être une application d'interprétation de commande, capable de piloter d'autres applications telles que :

- la consultation d'un document multimédia (comportant des rubriques par chapitres), ou
- le lancement d'appels téléphoniques, ou
- l'envoi de messages, ou
- une programmation d'assistance à la navigation,

ou autres.

En particulier, l'unité de traitement peut affecter des scores respectifs à une pluralité d'interprétations possible de la commande vocale prononcée pour :

- identifier au moins une interprétation de score le plus élevé,
- déterminer une rubrique du menu arborescent correspondant à cette interprétation, et
- proposer dans un signal de clarification des options à choisir et correspondant à des rubriques immédiatement en aval dans l'arborescence du menu.

A titre d'exemple purement illustratif, si l'utilisateur prononce « envoie un message », l'unité de traitement demande une clarification pour savoir s'il convient d'envoyer le message par courriel ou par messagerie SMS. Dans un autre exemple, si l'utilisateur prononce « appelle Monsieur Martin », l'unité de traitement demande une clarification pour savoir si l'appel doit être passé sur la ligne fixe ou sur le téléphone mobile du contact.

L'invention peut aller encore plus loin et dans une forme de réalisation complémentaire ou variante, l'unité de traitement peut affecter des scores respectifs à une pluralité d'interprétations de la commande vocale pour :
- identifier une ou plusieurs interprétations de scores plus élevés qu'un seuil choisi, et dans le cas d'une pluralité d'interprétations de la commande dépassant ledit seuil et correspondant à des rubriques de menu sur des branches séparées de l'arborescence,
- proposer dans un signal de clarification des options à choisir et correspondant auxdites rubriques.

Les branches séparées peuvent correspondre par exemple à des applications distinctes à lancer, mais néanmoins avec suffisamment d'informations dans la commande prononcée pour pouvoir exécuter effectivement chaque application.

Ainsi, par exemple si l'utilisateur prononce « je veux la maison », l'unité de traitement demande une clarification pour savoir si l'utilisateur souhaite passer un appel vers la maison (auquel cas il s'agit d'exécuter une application d'appel téléphonique vers le numéro du domicile utilisateur) ou s'il souhaite lancer l'application d'assistance à la navigation pour afficher le trajet vers son domicile.

Bien entendu, d'autres exemples sont possibles. Ainsi, les branches séparées peuvent correspondre par exemple à des chapitres séparés d'une même application, comme par exemple la consultation d'un document multimédia (par exemple un catalogue de matériel

automobile), chaque chapitre comportant lui-même des sous-chapitres. Dans cet exemple, chaque chapitre du catalogue référence tous les matériels d'une même marque automobile, et chaque sous-chapitre référence un type de matériel particulier de cette même marque (le système d'embrayage dans un sous-chapitre, le système de freins dans un autre sous-chapitre, etc.). Ainsi, dans ce cas, si l'utilisateur prononce « montre-moi les nouveaux systèmes d'embrayage pour boites automatiques, sortis en 2014 », l'unité de traitement note que la commande vocale est incomplète et demande des précisions, par exemple, sur la marque du constructeur (par exemple sous la forme de plusieurs choix possibles de constructeurs référencés par des numéros de 1 à N). Néanmoins, l'unité de traitement, après obtention de l'information quant à la marque souhaitée, oriente directement l'utilisateur vers le sous-chapitre des systèmes d'embrayage de cette marque.

Ainsi, on comprendra que cette forme de réalisation permet de croiser les différentes rubriques du menu arborescent avec les différentes interprétations possibles de la commande.

Ici, les scores respectifs des interprétations possibles peuvent être des scores de pertinence de la commande prononcée par rapport aux rubriques du menu arborescent, ou une pondération entre la pertinence et l'intelligibilité des mots prononcés (dans le cas où un doute subsiste quant aux mots effectivement prononcés).

Ainsi, par exemple, si l'utilisateur prononce « appelle chez ma mère » et que la qualité de l'acquisition sonore est médiocre pour la reconnaissance vocale, l'unité de traitement affecte des scores respectifs aux interprétations « Appelle Monsieur Mamère » (un contact existant de l'utilisateur) et « Appelle Maman » (autre contact existant de l'utilisateur). Dans ce cas, la demande de clarification consiste à solliciter l'utilisateur pour déterminer quel contact appeler parmi « Monsieur Mamère » et « Maman ».

Bien entendu, cette réalisation n'exclue pas d'affecter un score à chaque mot de la commande prononcée.

Ainsi, dans une réalisation complémentaire ou alternative, l'unité de traitement peut affecter des scores respectifs à une pluralité de mots de la commande vocale pour :

- identifier un ou plusieurs mots de scores plus élevés qu'un seuil choisi, et dans le cas d'une pluralité de mots dans la commande dépassant ledit seuil et correspondant à des rubriques de menu sur une même branche de l'arborescence,
- exécuter l'application jusqu'à la rubrique correspondante la plus avale de ladite même branche.

Il s'agit alors notamment du cas positif d'une commande éventuellement complète et, en tout cas bien interprétée en une fois par l'unité de traitement, sans nécessité de guider l'utilisateur jusqu'à la rubrique la plus avale précitée.

Dans une forme de réalisation complémentaire ou alternative, le score précité peut être fonction d'un contexte d'utilisation, en comparaison d'un historique d'interprétations validées (et/ou non-valides) de commandes précédentes en correspondance de contextes respectifs.

Ainsi, dans le même exemple précédent où l'utilisateur prononce « appelle chez ma mère » et que la qualité de l'acquisition sonore est médiocre, l'unité de traitement affecte des scores respectifs aux interprétations « Appelle Monsieur Mamère » (un contact existant de l'utilisateur) et « Appelle Maman » (autre contact existant de l'utilisateur). Dans ce cas par exemple, si l'utilisateur n'est plus en contact avec Monsieur Mamère et qu'il a l'habitude d'appeler sa mère toutes les semaines à une heure précise qui correspond à l'heure courante, l'unité de traitement identifie que le contact demandé est maman et lance l'appel vers ce contact sans nécessiter alors de clarification supplémentaire. On comprendra ainsi que cette forme de réalisation permet un apprentissage de l'intention-même de l'utilisateur.

L'invention propose alors une guidance efficace de l'utilisateur par rapport plusieurs choix possibles lorsque sa commande vocale est incomplète ou imprécise. Elle propose des scenarii de contournements adaptatifs pour obtenir de l'utilisateur les informations manquantes par des questions ou des options adaptées. Le critère « d'adaptivité » se base sur les scores précités d'intelligibilité de la commande. Typiquement, plus un score de commande est faible, plus le nombre d'options pour compléter la commande correspondante devrait être élevé.

Le principe de cette guidance peut s'appliquer aussi, dans une forme de réalisation alternative ou complémentaire, à l'acquisition-même du signal acoustique qu'émet l'utilisateur lorsqu'il prononce sa commande.

Dans cette forme de réalisation, au moins un microphone capte un signal de commande vocale d'un utilisateur. L'unité de traitement détermine un niveau d'intelligibilité du signal en particulier pour interpréter la commande vocale et, en cas d'inintelligibilité du signal lui-même, transmet à une interface homme/machine un signal de demande de répétition pour faire répéter par l'utilisateur une partie au moins de la commande vocale. Dans cette forme de réalisation, le procédé comporte en particulier les étapes dans lesquelles:

- on prévoit un jeu de paramètres du signal de demande de répétition, propre à au moins un élément parmi une forme, une temporalité et une vitesse de prononciation de la commande vocale par l'utilisateur,
- l'unité de traitement évalue un score d'intelligibilité d'un signal de parole couramment prononcé, et
- l'unité de traitement affecte des valeurs à ces paramètres en fonction du score d'intelligibilité précité pour construire le signal de demande de répétition.

Par exemple, le signal de demande de répétition peut comporter au moins une demande de répétition parmi :

- une répétition par l'utilisateur à un instant paramétré, choisi en fonction dudit score,
- une répétition à une distance paramétrée du microphone, choisie en fonction dudit score, et
- une répétition par l'utilisateur avec une prononciation plus lente.

Le score d'intelligibilité du signal peut être évalué au moins en fonction d'un élément parmi un bruit ambiant (ou une variation de bruit ambiant) et un historique d'échecs d'interprétation de commandes vocales précédentes.

Ainsi, dans cette forme de réalisation, le principe de guidance de l'utilisateur est mis en œuvre dans le cadre d'une application d'acquisition du signal de commande vocale. Les améliorations apportées se sont révélées significatives. Ce mode de réalisation, particulièrement avantageux, peut faire l'objet d'une protection en soi (séparément d'une guidance dans l'interprétation de la commande après son acquisition sonore).

La présente invention vise aussi un programme informatique comportant des instructions pour la mise en œuvre du procédé ci-avant, lorsque ce programme est exécuté par un processeur.

Un exemple d'ordinogramme de l'algorithme général d'un tel programme est illustré sur la figure 2 commentée plus loin.

En outre, la présente invention vise aussi un dispositif informatique de reconnaissance vocale, comportant une unité de traitement pour la mise en œuvre du procédé ci-avant.

D'autres caractéristiques et avantages de l'invention apparaîtront à l'examen de la description détaillée ci-après, et des dessins annexés sur lesquels :

- la figure 1 illustre un exemple de réalisation d'un dispositif au sens de l'invention ;
- la figure 2 illustre les principales étapes d'un procédé selon un exemple de réalisation de l'invention ;
- la figure 3 détaille à titre d'exemple les principales étapes d'une attribution de score d'interprétation d'une commande selon l'étape S3 de la figure 2, dans un mode de réalisation particulier ;
- la figure 4 détaille à titre d'exemple les principales étapes d'une amélioration de l'acquisition sonore selon l'étape S2 de la figure 2, dans un mode de réalisation particulier ;
- la figure 5 illustre un exemple d'arborescence d'un menu d'application d'interprétation, dans une forme de réalisation d'un programme au sens de l'invention.

On se réfère à la figure 1 sur laquelle on a représenté un dispositif DIS selon l'invention comportant une interface d'entrée INE pour être connecté à au moins un microphone MIC (directement ou via un réseau local RES, par exemple au sein d'un véhicule automobile). Le dispositif DIS comprend en outre une unité de traitement incluant un processeur PROC et une mémoire de travail MEM (ou tout autre module informatique à processeur intégré, tel qu'un ASIC ou autre). Le dispositif comporte en outre une interface de sortie INS pour transmettre des instructions d'affichage à un écran SCR et/ou des haut-parleurs HP d'une interface homme/machine proposant à un utilisateur des options à choisir. Le dispositif DIS peut en outre piloter le lancement d'applications spécifiques APP comme le lancement d'un appel téléphonique, d'une assistance à la navigation vers un lieu précisé (NAV), ou encore d'un

message à communiquer à un contact (MES). Ainsi, le dispositif DIS peut piloter d'autres équipements tels qu'un téléphone de type Smartphone par exemple, ou encore un appareil de navigation, ou autres, ou encore piloter l'écran SCR lui-même pour afficher un contenu multimédia (un catalogue, des informations d'actualité, etc.).

Ainsi, un utilisateur parle dans le microphone MIC et le dispositif DIS interprète la phrase prononcée comme une commande vocale afin de lancer une application APP correspondante.

Pour aider le dispositif dans la reconnaissance vocale de la phrase prononcée, une base de données DB d'un historique des commandes vocales validées précédemment est stockée dans une mémoire durable connectée au dispositif DIS afin qu'il puisse s'y référer.

En référence maintenant à la figure 2, un utilisateur prononce une phrase à l'étape S1. À l'étape S2, une estimation de vraisemblance des mots prononcés est effectuée et, si cette estimation dépasse un seuil par exemple, une estimation des différentes interprétations possibles de la phrase est ensuite menée à l'étape S3. Des détails de ces étapes S2 et S3 sont présentés plus loin en référence aux figures respectives 4 et 3.

À l'étape subséquente S4, l'unité de traitement détermine s'il existe une unique interprétation possible INT dont l'estimation de vraisemblance dépasse un seuil THR prédéterminé. Ce seuil peut être fonction de plusieurs critères, et notamment de la qualité de l'acquisition sonore d'après le score calculé à l'étape S2. L'estimation de la vraisemblance d'une interprétation INT s'effectue en déterminant s'il existe dans la phrase prononcée un ou plusieurs mots d'estimation élevée à l'étape S2 et qui correspondent tous à une même rubrique dans un menu arborescent présenté plus loin en référence à la figure 5. Par ailleurs, l'unité de traitement peut se référer à un historique des commandes vocales validées dans la base de données DB pour augmenter le score de cette interprétation si le nombre d'occurrences d'une commande similaire est élevé dans des conditions similaires d'utilisation (par exemple en fonction de l'horodate courante, de la localisation de l'utilisateur, etc.).

S'il existe une unique interprétation dont l'estimation dépasse un seuil (flèche OK en sortie du test S4), à l'étape suivante S5, l'unité de traitement vérifie si cette interprétation correspond à une commande vocale complète permettant de lancer directement une application : par

exemple « appelle Gérard Dupont sur son téléphone mobile » (flèche OK en sortie du test S5), auquel cas, à l'étape S6, l'application APP correspondante peut être directement exécutée. Le procédé de reconnaissance vocale s'arrête à l'étape S7 et à l'étape suivante S8, des informations supplémentaires peuvent être archivées dans la base DB, telles que la phrase prononcée PRON, l'interprétation INT qui en est faite (éventuellement validée par l'utilisateur comme on le verra plus loin), l'horodate courante HRD, la localisation courante LOC de l'utilisateur, et autres.

À l'étape S9, dans le cas où la phrase prononcée ne correspond pas à une commande vocale complète (flèche KO en sortie du test S5), l'unité de traitement consulte la ou les rubriques manquantes dans la commande vocale pour qu'elle soit complète. A l'étape suivante S10, l'unité de traitement anime l'interface homme-machine IHM (constituée de l'écran SCR et des haut-parleurs HP) pour proposer des options OPT11, OPT12 à l'utilisateur permettant de renseigner la ou les rubriques manquantes (par exemple par affichage de ces options sur l'écran, en attendant une sélection de l'utilisateur par commande tactile ou via le microphone). Par exemple : « appelle Gérard Dupont » peut être une commande incomplète et l'unité de traitement peut proposer les options : « sur son mobile » et « ou sur sa ligne fixe ». A l'étape suivante S11, l'utilisateur rentre l'une des options proposées OPT11 et le traitement peut reprendre à l'étape S5 pour vérifier que la commande vocale est maintenant complète.

En réalité, les étapes S9 à S11 sont optionnelles car elles correspondent à des situations où, du fait de l'incomplétude de la commande, plusieurs interprétations sont possibles comme : « appelle Gérard Dupont sur son mobile » ou « appelle Gérard Dupont sur sa ligne fixe ». Ces étapes S9 à S11 peuvent être mises en œuvre à des fins d'apprentissage à l'étape S8, une fois la commande complète et bien interprétée à l'étape S6. Toutefois, cette réalisation sollicite l'intervention de l'utilisateur et, dans une variante illustrée aussi sur la figure 2, on vérifie avant de solliciter l'utilisateur que plusieurs interprétations possibles d'une commande dépassent toutes un seuil de vraisemblance THR (flèche KO en sortie du test S4). Par exemple si l'utilisateur a dans son carnet d'adresse « Gérard Dupont » (contact appelé fréquemment) et « Michel Dupont » (non contacté depuis plusieurs années), et que la phrase prononcée est « appelle Dupont sur son téléphone mobile », l'interrogation de la base de données DB peut donner :

- un faible score pour l'interprétation « appelle Michel Dupont sur son téléphone mobile », inférieur au seuil THR ; et
- un grand score pour l'interprétation « appelle Gérard Dupont sur son téléphone mobile », supérieur au seuil THR.

Dans ce cas, l'application de lancement d'appel est exécutée vers le téléphone mobile de Gérard Dupont, sans solliciter l'intervention de l'utilisateur.

Sinon (dans le cas où les scores sont supérieurs au seuil THR), l'unité de traitement peut animer l'interface IHM à l'étape S12 pour afficher les options OPT1, OPT2, … afin que l'utilisateur choisisse la bonne option (par exemple Michel ou Gérard Dupont). Bien entendu, si aucune des options ne correspond à un choix souhaité par l'utilisateur, l'interface IHM offre une possibilité RET de retourner à d'autres options amont à choisir (par exemple : « souhaitez-vous bien appeler Dupont ? », et, si la réponse est non : « qui souhaitez-vous appeler ? »), ou dans le pire cas à une nouvelle acquisition sonore. Si l'une des options proposées (par exemple OPT2) est retenue par l'utilisateur à l'étape S13, l'unité de traitement vérifie à nouveau s'il n'existe plus qu'une interprétation possible à l'étape S4 et, le cas échéant, vérifie encore la complétude de la commande vocale à l'étape S5. Éventuellement, le seuil THR peut être augmenté après chaque itération de l'étape S13 car, en principe, une fois qu'une option a été renseignée, la vraisemblance d'une interprétation particulière de la commande devrait augmenter significativement et d'autres interprétations devraient être de scores plus faibles.

Il apparaît d'après l'exemple ci-avant du lancement d'un appel téléphonique que l'interprétation d'une commande suit la logique d'une arborescence visant :

- le type d'application à lancer (appel téléphonique, ou envoi de message par exemple),
- puis le contact du carnet d'adresse à appeler,
- puis l'appareil de destination (ligne fixe ou téléphone mobile).

Ici, l'application peut interroger le carnet électronique de contacts de l'utilisateur et déterminer si la ligne fixe et le téléphone mobile sont bien des données renseignées dans le carnet de contacts, avant de proposer un choix entre les deux options. On comprendra ainsi que l'application se réfère à un contexte informatique (notamment de données stockées en relation avec diverses autres applications) avant de proposer des options.

On a illustré un exemple d'une telle arborescence sur la figure 5. Ainsi, le nœud racine d'une interprétation de commande INTC peut se décomposer en plusieurs branches correspondant à des applications différentes, telles que :

- le lancement d'un appel téléphonique TEL,
- la consultation d'un document multimédia MM,
- le lancement d'une application de messagerie MES, et dans ce cas, de type SMS sur un sous-branche ou de type messagerie instantanée (ou « chat ») sur un autre sous-branche WAP,
- ou encore une application d'assistance à la navigation (non représentée), ou autres.

Chacune de ces branches ou sous-branches se décompose encore en sous-branches liées à différents contacts CTC d'un carnet d'adresse, et encore ensuite en deux sous-branches dans le cas d'un contact renseigné par son numéro de mobile MOB et son numéro de fixe FIX. Les feuilles de cette arborescence correspondent à la commande vocale complète permettant d'exécuter une application sans nécessiter d'intervention supplémentaire de l'utilisateur, comme par exemple : « appeler Michel Dupont sur son terminal mobile » (feuille AMM de l'arborescence).

Ainsi, on comprendra que l'application interactive d'interprétation de commande vocale permet de croiser les interprétations d'une phrase prononcée avec différentes branches de cette arborescence. L'utilisateur est alors guidé pour éliminer progressivement les branches non pertinentes, puis éventuellement, dans une même branche, guidé encore pour préciser la commande pour qu'elle soit complète.

Par exemple, si l'utilisateur prononce la phrase : « laisse un message sur le mobile de Dupont », des interprétations possibles sont :

- appelle le téléphone mobile de Dupont : CR1 (de sorte que l'utilisateur laisse un message vocal),
- ouvre l'application de messagerie SMS vers le téléphone mobile de Dupont pour entrer un message : CR2 ;
- ouvre l'application de messagerie instantanée vers le téléphone mobile de Dupont pour entrer un message : CR3.

Dans ce cas, l'unité de traitement élimine progressivement les branches non pertinentes et anime l'interface homme-machine pour proposer à l'utilisateur de sélectionner l'émission d'un

appel téléphonique ou d'un message. Si l'émission d'un message est choisie, l'interface homme-machine demande s'il s'agit d'un message SMS ou d'un message instantané. Ensuite, il peut être demandé à l'utilisateur de préciser le contact dans la branche de l'application de messagerie SMS (« Michel ou Gérard Dupont ?»). Bien entendu, alternativement, cette précision peut être apportée avant de déterminer le type d'application à lancer, sans perte de généralité. Par ailleurs, comme indiqué plus haut, la possibilité d'envoyer un message de type courriel par exemple est exclue d'emblée si le carnet de contacts de l'utilisateur n'a aucune donnée renseignée sur une éventuelle adresse email.

On a illustré aussi sur la figure 5 l'exemple la consultation d'un document multimédia MM (par exemple un catalogue de matériel automobile), comportant plusieurs chapitres M1, M2,..., MN et chaque chapitre M1 comportant lui-même des sous-chapitres SE1, ..., SF1. Dans cet exemple, chaque chapitre CH du catalogue référence tous les matériels d'une même marque automobile M1, M2, ..., MN, et chaque sous-chapitre SCH référence un type de matériel particulier de cette même marque (le système d'embrayage SE2 dans un sous-chapitre, le système de freins SF2 dans un autre sous-chapitre, etc., pour la marque M2 par exemple). Ainsi, dans ce cas, si l'utilisateur prononce « montre-moi les systèmes d'embrayage pour boites automatiques » sans préciser la marque du constructeur, l'unité de traitement note que la commande vocale est incomplète car elle mène à plusieurs choix C1 (sur la sous-branche SE1), C2 (sur la sous-branche SE2), etc., correspondant à des systèmes d'embrayage pour boite automatique proposés par différentes marques. L'interface IHM demande alors des précisions, par exemple, sur la marque du constructeur (par exemple sous la forme de plusieurs choix possibles de constructeurs référencés par des numéros de 1 à N sur l'écran de l'interface IHM). Néanmoins, l'unité de traitement, après obtention de l'information quant à la marque souhaitée (M2), oriente directement l'utilisateur vers le sous-chapitre SE2 des systèmes d'embrayage de cette marque.

On se réfère maintenant à la figure 3 pour décrire une réalisation d'affectation de scores de vraisemblance à des interprétations d'une phrase prononcée. Le traitement commence après la phrase prononcée à l'étape S1 de la figure 2, et après une évaluation des scores acoustiques des mots identifiés à l'étape S2 de la figure 2 qui sera commentée en détails plus loin en référence à la figure 4. Les scores acoustiques SCMi associés à chaque mot Mi étant estimés, le traitement se poursuit à l'étape S31 de la figure 3 par une comparaison de chaque score à un

seuil prédéterminé THR' (bien entendu, différent du seuil THR d'interprétation, précité). Si aucun des scores ne dépasse ce seuil (flèche KO en sortie du test S31), alors la phrase prononcée est simplement inintelligible et l'interface IHM invite l'utilisateur à la prononcer à nouveau, en modifiant par exemple les paramètres d'acquisition sonore comme on le verra plus loin en référence à la figure 4. S'il existe au moins un mot Mi dont le score est supérieur au seuil THR' (flèches OK en sortie des tests S31 et S32), on détermine à l'étape S33 s'il existe au moins une rubrique R du menu arborescent, associée à ce mot Mi (se présentant typiquement comme un mot-clé pour une application (comme « Appelle » ou « Message») ou pour un sous-branche d'une application (« Dupont » identifié comme un contact mais sans précision de l'application correspondante à lancer). Si aucune rubrique ne peut être associée à ce mot Mi (flèche KO en sortie du test S33), il peut être prévu d'inviter l'utilisateur à prononcer à nouveau sa commande, comme indiqué précédemment. En revanche, s'il existe bien une rubrique associée au mot Mi (flèche OK en sortie du test S33), alors il convient de déterminer un degré de pertinence de cette rubrique par rapport à un contexte courant, à l'étape S36. Par exemple, si le mot Mi identifié est « Athènes », alors que l'utilisateur ne s'y est jamais rendu et qu'il est localisé actuellement, ainsi que de façon générale (d'après l'historique de la base BD), en région parisienne, alors le score de pertinence de la seule rubrique associée à ce mot et correspondant à l'interprétation de la commande vocale : « assiste ma navigation jusqu'à Athènes », est plus faible que le seuil de vraisemblance d'interprétation THR. Dans ce cas, il n'existe pas d'interprétation possible d'une commande associée à ce mot (flèche OK en sortie du test S36) et l'utilisateur est invité à reformuler sa commande. En revanche, s'il existe au moins une interprétation dont la vraisemblance dépasse le seuil THR, comme par exemple « assiste ma navigation jusqu'à la rue d'Athènes » (flèche KO en sortie du test S36), alors le procédé se poursuit par l'étape S4 de la figure 2, dans laquelle il est déterminé en outre s'il existe plusieurs interprétations vraisemblables (comme par exemple : « assiste ma navigation jusqu'à la rue d'Athènes à Paris », ou « assiste ma navigation jusqu'à la rue d'Athènes à Lyon », etc., auquel cas l'interface IHM peut être animée pour demander à l'utilisateur de compléter sa commande, comme présenté ci-avant).

S'il existe plusieurs mots M1…Mj dont les scores acoustiques sont tous supérieurs au seuil THR' (flèche KO en sortie du test S32) et que tous ces mots concernent une rubrique unique (« Affiche » « catalogue » « automobile ») (flèche OK en sortie du test S34), cette situation correspond au cas favorable où la phrase prononcée est cohérente et, au test S36, son

estimation de vraisemblance devrait dépasser le seuil THR. Si tel est le cas (flèche KO en sortie du test S36), alors la commande est complète ou ne nécessite qu'un nombre limité d'interactions avec l'utilisateur pour être complétée après le test S5.

En revanche, si plusieurs rubriques R1, R2, … peuvent être associées à l'ensemble de mots M1 … Mj de scores acoustiques supérieurs au seuil THR' (flèche KO en sortie du test S34), alors on évalue à l'étape S35 des scores respectifs de vraisemblance d'interprétation de chaque rubrique R1, R2, … Ainsi, il est possible, à cette étape, de se référer à la base de données DB pour déterminer une cohérence de l'interprétation (« assiste ma navigation jusqu'à Athènes ») par rapport à un contexte courant (localisation en région parisienne), en comparaison d'un historique (aucun voyage à Athènes déjà identifié) et attribuer un score correspondant (faible dans cet exemple).

Néanmoins, les scores acoustiques des mots peuvent intervenir encore à l'étape S35, pour l'estimation des vraisemblances. Par exemple, si les mots dont les scores dépassent le seuil THR' sont :
« Appelle » ; « chez » , « Michel » ; « ma mère » , « Mamère »,
alors les interprétations possibles sont :
« appelle chez ma mère » et « appelle Michel Mamère ».
Le typiquement ici, si le score acoustiques de « Michel », bien que supérieure au seuil THR', est inférieur à celui des autres mots, l'interprétation « appelle chez ma mère » aura un meilleur score. Ainsi, on comprendra que les scores de vraisemblance des interprétations peuvent résulter d'une pondération entre le score acoustique des mots, le nombre de mots significatifs associés à une même rubrique, la comparaison d'un contexte courant à un historique après consultation de la base DB.

Dans ce cas, après l'étape S35, on détermine s'il existe au moins une interprétation dont le score de vraisemblance dépasse le seuil THR à l'étape S36. Si tel n'est pas le cas (flèche KO en sortie du test S36), aucune interprétation de l'ensemble des mots prononcés n'est plausible et il est demandé à l'utilisateur de prononcer sa commande à nouveau. Sinon, on détermine au test S4 de la figure 2 si plusieurs interprétations ont un score de vraisemblance supérieur au seuil THR (flèche KO en sortie du test S4) et ainsi, comme indiqué ci-avant, à l'étape S12, l'utilisateur est guidé pour déterminer la commande qu'il entendait prononcer (appeler :

« Maman » ou « Michel Mamère » ?). Dans cet exemple, alors que la commande initialement prononcée était particulièrement équivoque, l'utilisateur a simplement ici à entrer un choix parmi deux options, alors que les assistants personnels intelligents existants auraient demandé une reformulation complète de la phrase pour discriminer entre les prononciations de « Michel » et de « chez ». On comprendra ainsi que cette guidance intelligente est un atout fort de la présente invention.

Le principe de cette guidance peut être appliqué aussi à l'acquisition-même du signal vocal que prononce utilisateur. Ainsi, en référence à la figure 2, si les scores acoustiques estimés à l'étape S2 ne révèlent pas de mots significatifs (de score supérieur à THR'), l'utilisateur est guidé dans une nouvelle prononciation de sa commande, en référence à la figure 4. La première étape S41 de ce traitement consiste, dans un exemple de réalisation, à mémoriser le nombre N d'échecs précédents Ech. À l'étape S42, le traitement calcule un rapport signal à bruit RSB du signal acoustique prononcé par l'utilisateur.

Par exemple, si le nombre d'échecs N est supérieur à un seuil THR1 alors que le rapport signal à bruit reste supérieur à un deuxième seuil THR2, une telle situation signifie que l'utilisateur, malgré de bonnes conditions d'acquisition sonore, ne prononce pas de façon intelligible sa commande (flèche OK en sortie du test S43), et il lui est demandé, dans une étape ultérieure S44 de prononcer plus lentement sa commande.

Hormis cette situation, il est possible que le rapport signal à bruit RSB, en moyenne, soit faible et, dans ce cas, il est procédé à une analyse de la variation dans le temps du rapport signal à bruit à l'étape S45. Si cette variation est supérieure à un seuil THR3 (flèche OK en sortie du test S45), cette situation signifie que des bruits sont présents par intermittence et il est procédé à l'étape S47 à une temporisation pendant laquelle l'utilisateur n'est pas invité à prononcer sa commande, et ce jusqu'à ce que le bruit ambiant soit inférieur à un seuil THR4 (flèche OK en sortie du test S48). À cet instant alors, l'interface IHM est animée pour inviter l'utilisateur à prononcer à nouveau sa commande à l'étape S49.

Si la variation temporelle du rapport signal à bruit reste inférieure au seuil THR3 (flèche KO en sortie du test S45), cette situation signifie qu'il n'y a pas d'amélioration du rapport signal à bruit à espérer et, dans ce cas, il est demandé à l'utilisateur de s'approcher du microphone pour

prononcer à nouveau sa commande, à l'étape S46. Cette précaution contribue à améliorer le rapport signal à bruit.

Bien entendu, la présente invention ne se limite pas aux formes de réalisation décrites ci-avant à titre d'exemple ; elle s'étend à d'autres variantes.

Typiquement, les comparaisons à des seuils, présentées ci-avant, correspondent à des exemples de réalisation pratique, mais leur description ne limite en rien l'interprétation générale des principales étapes de l'invention, consistant en particulier à guider l'utilisateur par des choix d'options judicieuses.

De même, la présentation du rapport signal à bruit en référence à la figure 4 pour le critère de la qualité de l'acquisition sonore à évaluer est un exemple de réalisation. D'autres critères d'évaluation sont possibles (par exemple, le niveau de captation (mesuré en décibels), et/ou son évolution temporelle).

Revendications

1. Procédé mis en œuvre par des moyens informatiques de reconnaissance vocale, dans lequel une unité de traitement détermine un niveau d'intelligibilité d'un commande vocale prononcée par un utilisateur et, en cas d'inintelligibilité de la commande, transmet à une interface homme/machine un signal de demande de clarification par l'utilisateur d'une partie au moins de la commande vocale,

le procédé comportant les étapes dans lesquelles:

- l'unité de traitement évalue un score d'intelligibilité de la commande vocale, et
- l'unité de traitement transmet à l'interface homme/machine un signal de demande de clarification en fonction dudit score d'intelligibilité.

2. Procédé selon la revendication 1, dans lequel l'unité de traitement exécute une application comportant un menu à rubriques arborescentes, et dans lequel l'unité de traitement affecte des scores respectifs à une pluralité d'interprétations de la commande vocale pour :

- identifier au moins une interprétation de score le plus élevé,
- déterminer une rubrique du menu correspondant à cette interprétation,
- proposer dans un signal de clarification des options à choisir et correspondant à des rubriques immédiatement avales dans l'arborescence du menu.

3. Procédé selon l'une des revendications précédentes, dans lequel l'unité de traitement exécute une application comportant un menu à rubriques arborescentes, et dans lequel l'unité de traitement affecte des scores respectifs à une pluralité d'interprétations de la commande vocale pour :

- identifier une ou plusieurs interprétations de scores plus élevés qu'un seuil choisi, et dans le cas d'une pluralité d'interprétations de la commande dépassant ledit seuil et correspondant à des rubriques de menu sur des branches séparées de l'arborescence,
- proposer dans un signal de clarification des options à choisir et correspondant auxdites rubriques.

4. Procédé selon l'une des revendications précédentes, dans lequel l'unité de traitement exécute une application comportant un menu à rubriques arborescentes, et dans lequel l'unité de traitement affecte des scores respectifs à une pluralité de mots de la commande vocale pour :
- identifier un ou plusieurs mots de scores plus élevés qu'un seuil choisi, et dans le cas d'une pluralité de mots dans la commande dépassant ledit seuil et correspondant à des rubriques de menu sur une même branche de l'arborescence,
- exécuter l'application jusqu'à la rubrique correspondante la plus avale de ladite même branche.

5. Procédé selon l'une des revendications précédentes, dans lequel le score est fonction d'un contexte d'utilisation, en comparaison d'un historique d'interprétations validées de commandes précédentes en correspondance de contextes respectifs.

6. Procédé selon l'une des revendications précédentes, dans lequel au moins un microphone capte un signal de commande vocale d'un utilisateur et l'unité de traitement détermine un niveau d'intelligibilité du signal pour interpréter la commande vocale et, en cas d'inintelligibilité du signal, transmet à une interface homme/machine un signal de demande de répétition pour faire répéter par l'utilisateur une partie au moins de la commande vocale, le procédé comportant les étapes dans lesquelles:
- on prévoit un jeu de paramètres du signal de demande de répétition, propre à au moins un élément parmi une forme, une temporalité et une vitesse de prononciation de la commande vocale par l'utilisateur,
- l'unité de traitement évalue un score d'intelligibilité d'un signal de parole couramment prononcé, et
- l'unité de traitement affecte des valeurs auxdits paramètres en fonction dudit score d'intelligibilité pour construire le signal de demande de répétition.

7. Procédé selon la revendication 6, dans lequel le signal de demande de répétition comporte au moins une demande de répétition parmi :
- une répétition par l'utilisateur à un instant paramétré, choisi en fonction dudit score,
- une répétition à une distance paramétrée du microphone, choisie en fonction dudit score, et
- une répétition par l'utilisateur avec une prononciation plus lente.

8. Procédé selon l'une des revendications 6 et 7, dans lequel le score est évalué au moins en fonction d'un élément parmi un bruit ambiant et un historique d'échecs d'interprétation de commandes vocales précédentes.

9. Programme informatique caractérisé en ce qu'il comporte des instructions pour la mise en œuvre du procédé selon l'une des revendications 1 à 8, lorsque ce programme est exécuté par un processeur.

10. Dispositif informatique de reconnaissance vocale, comportant une unité de traitement pour la mise en œuvre du procédé selon l'une des revendications 1 à 8.

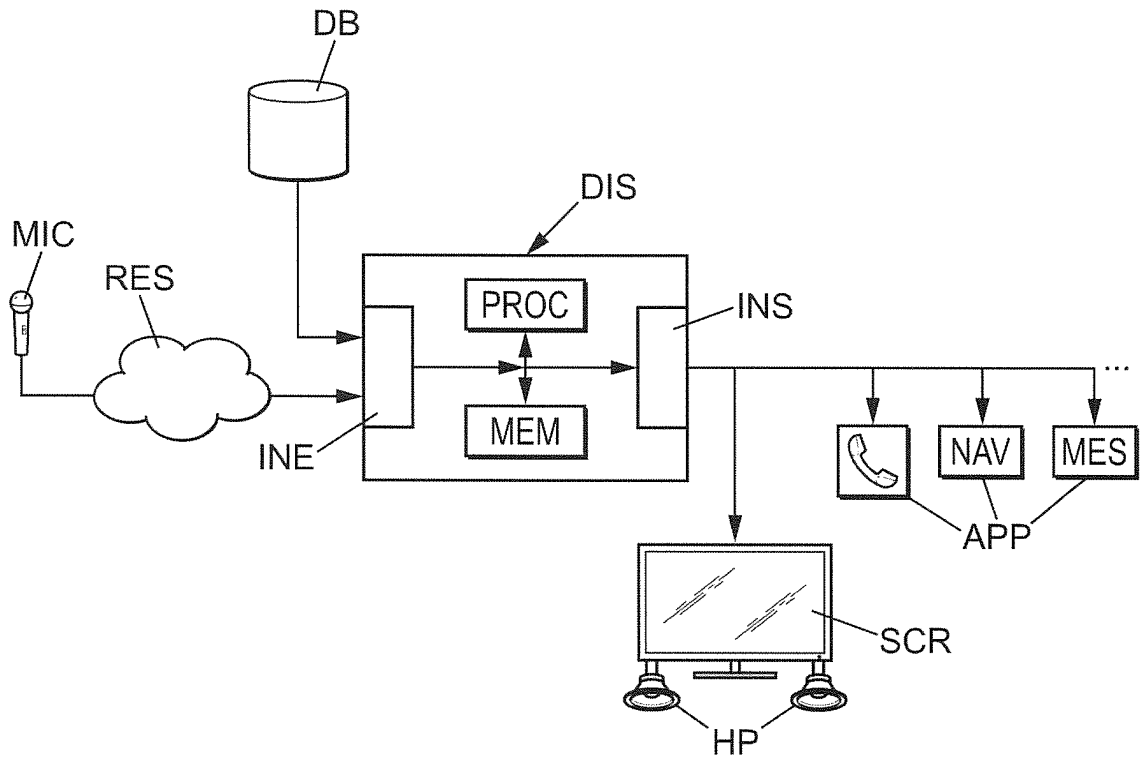

DB
DIS
MIC
RES
PROC
INS
MEM
INE
NAV
MES
APP
SCR
HP

FIG. 1

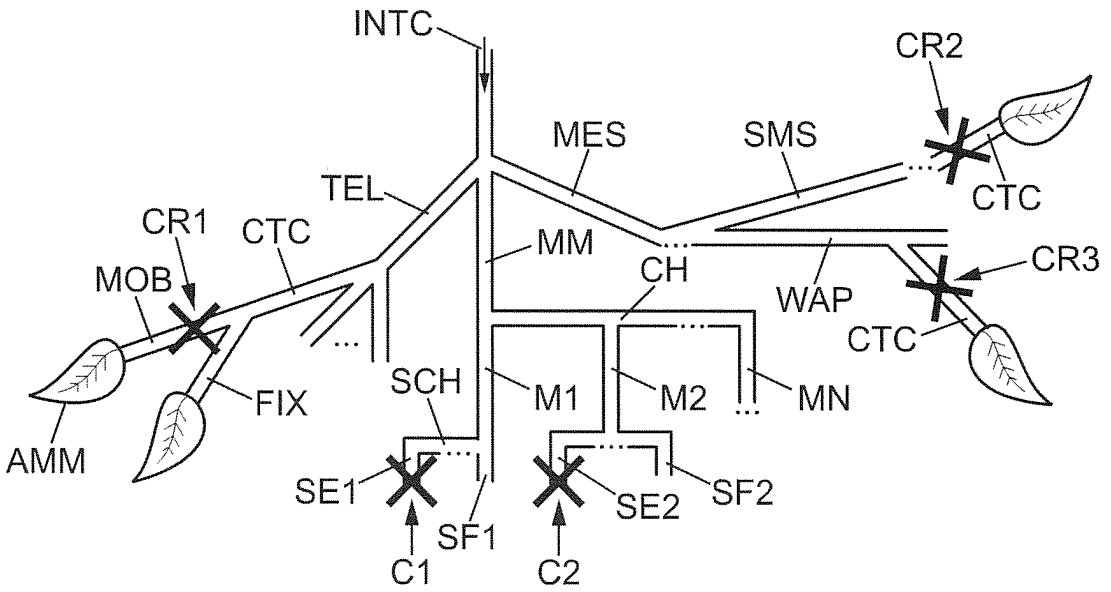

INTC
CR2
MES
SMS
CTC
TEL
CR1
CTC
MM
CR3
MOB
CH
WAP
CTC
SCH
M1
M2
MN
FIX
AMM
SE1
SE2
SF2
SF1
C1
C2

FIG. 5

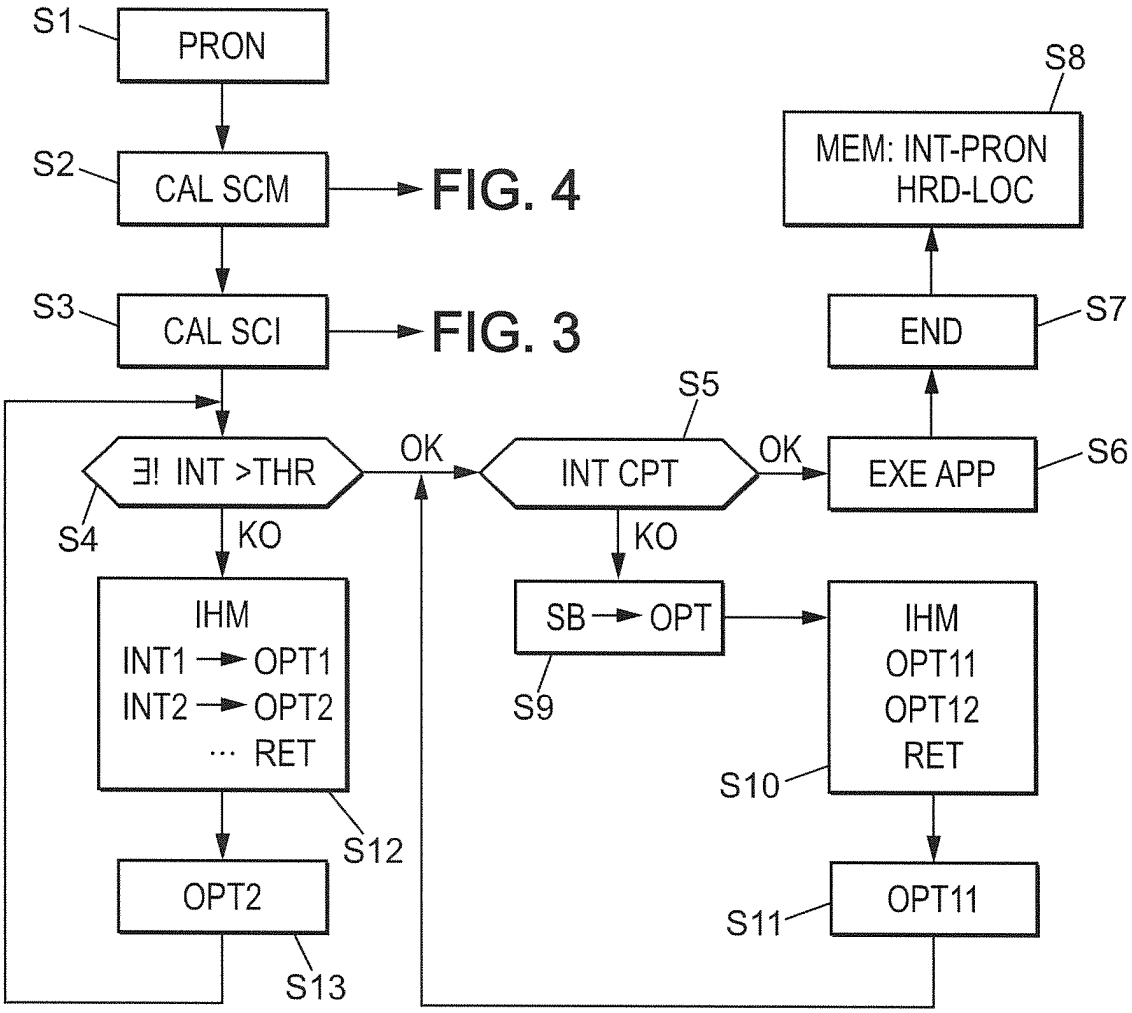

S1
PRON
S2
CAL SCM
FIG. 4
S3
CAL SCI
FIG. 3
∃! INT >THR
S4
OK
KO
INT CPT
S5
OK
KO
EXE APP
S6
END
S7
MEM: INT-PRON
HRD-LOC
S8
IHM
INT1 → OPT1
INT2 → OPT2
··· RET
S12
OPT2
S13
SB → OPT
S9
IHM
OPT11
OPT12
RET
S10
OPT11
S11

FIG. 2

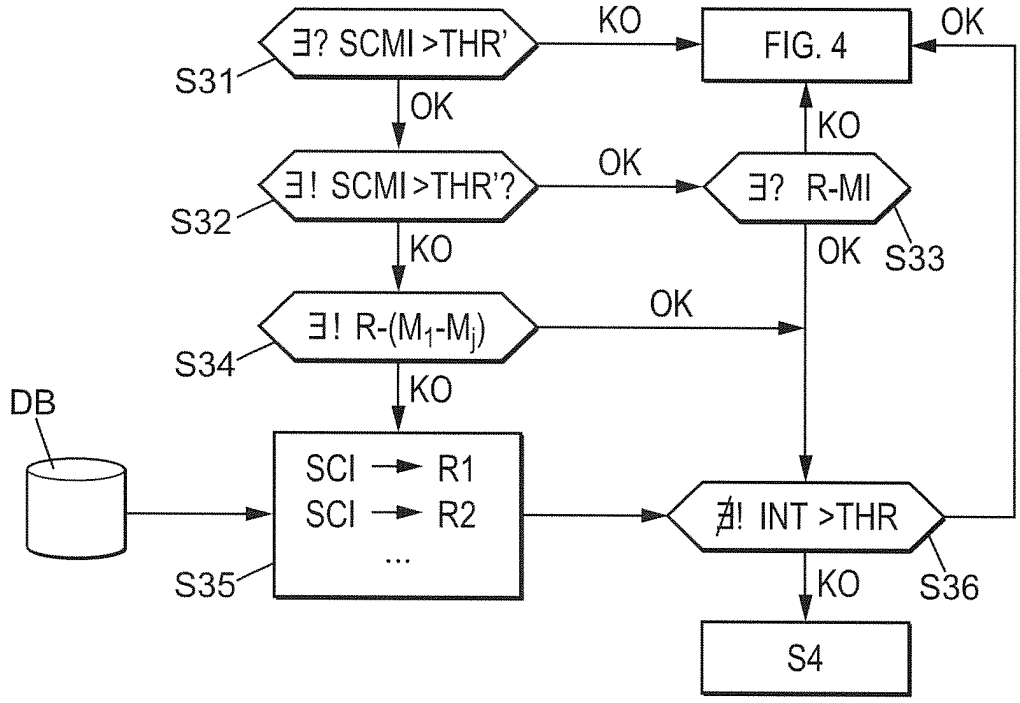

∃? SCMI >THR'
KO
FIG. 4
OK
S31
OK
KO
∃! SCMI >THR'?
OK
∃? R-MI
S32
KO
OK
S33
∃! R-(M1-Mj)
OK
S34
KO
DB
SCI → R1
SCI → R2
...
∄! INT >THR
S35
KO
S36
S4

FIG. 3

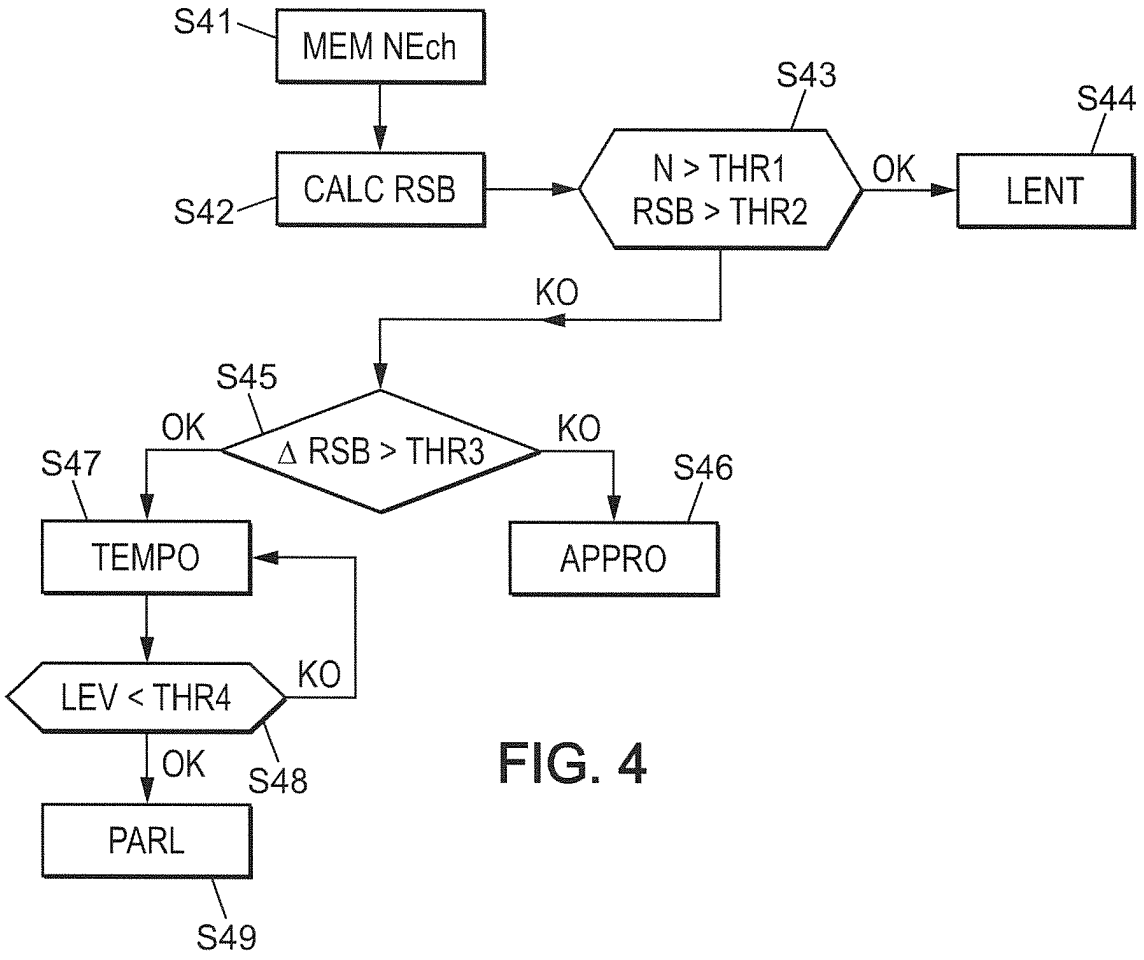

S41
MEM NEch
S42
CALC RSB
S43
N > THR1
RSB > THR2
OK
S44
LENT
KO
S45
OK
Δ RSB > THR3
KO
S47
S46
TEMPO
APPRO
LEV < THR4
KO
OK
S48
PARL
S49

FIG. 4

RÉPUBLIQUE FRANÇAISE

INSTITUT NATIONAL DE LA PROPRIÉTÉ INDUSTRIELLE

RAPPORT DE RECHERCHE PRÉLIMINAIRE

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

N° d'enregistrement national

FA 801080
FR 1459325

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 Y	EP 1 679 694 A1 (AT & T CORP [US]) 12 juillet 2006 (2006-07-12) * alinéa [[0026]]; figure 3 *	1,9,10 2-8	G10L15/22
Y A	US 6 044 347 A (ABELLA ALICIA [US] ET AL) 28 mars 2000 (2000-03-28) * colonne 7, ligne 53 - ligne 67 * * colonne 9, ligne 14 - ligne 67; figures 2,3 *	2-5 1,6-10	
Y A	US 2011/282669 A1 (MICHAELIS PAUL ROLLER [US]) 17 novembre 2011 (2011-11-17) * alinéa [[0016]] *	6-8 1,9,10	
X	Malte Gabsdil ET AL: "Combining Acoustic Confidence Scores with Deep Semantic Analysis for Clarification Dialogues", proceding of the 5th international workshop on computational semantics, 1 janvier 2003 (2003-01-01), XP55188800, Extrait de l'Internet: URL:http://www.let.rug.nl/bos/pubs/GabsdilBos2003IWCS.pdf [extrait le 2015-05-12] * page 4, ligne 1 - ligne 30 *	1,9,10	DOMAINES TECHNIQUES RECHERCHÉS (IPC) G10L
A	US 2012/053945 A1 (GUPTA RAKESH [US] ET AL) 1 mars 2012 (2012-03-01) * page 8, alinéa [0070] - alinéa [0073]; figure 7 *	1-10	

Date d'achèvement de la recherche	Examinateur
13 mai 2015	Perez Molina, E

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)

ANNEXE AU RAPPORT DE RECHERCHE PRÉLIMINAIRE RELATIF A LA DEMANDE DE BREVET FRANÇAIS NO. FR 1459325 FA 801080

La présente annexe indique les membres de la famille de brevets relatifs aux documents brevets cités dans le rapport de recherche préliminaire visé ci-dessus.
Les dits membres sont contenus au fichier informatique de l'Office européen des brevets à la date du 13-05-2015
Les renseignements fournis sont donnés à titre indicatif et n'engagent pas la responsabilité de l'Office européen des brevets, ni de l'Administration française

Document brevet cité au rapport de recherche	Date de publication	Membre(s) de la famille de brevet(s)	Date de publication
EP 1679694 A1	12-07-2006	CA 2531455 A1	05-07-2006
		DE 602006000090 T2	11-09-2008
		EP 1679694 A1	12-07-2006
		US 2006149544 A1	06-07-2006
US 6044347 A	28-03-2000	AUCUN	
US 2011282669 A1	17-11-2011	CN 102254556 A	23-11-2011
		GB 2480537 A	23-11-2011
		US 2011282669 A1	17-11-2011
US 2012053945 A1	01-03-2012	JP 5695199 B2	01-04-2015
		JP 2013542484 A	21-11-2013
		US 2012053945 A1	01-03-2012
		WO 2012030838 A1	08-03-2012

EPO FORM P0465