(19) RÉPUBLIQUE FRANÇAISE

INSTITUT NATIONAL DE LA PROPRIÉTÉ INDUSTRIELLE

COURBEVOIE

(11) **N° de publication :** ***3 059 191***

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

(21) **N° d'enregistrement national :** **16 61324**

(51) Int Cl[8] : ***H 04 R 5/033*** *(2017.01), H 04 R 3/00, G 10 K 11/16*

(12) DEMANDE DE BREVET D'INVENTION A1

(22) **Date de dépôt :** 21.11.16.

(30) **Priorité :**

(43) **Date de mise à la disposition du public de la demande :** 25.05.18 Bulletin 18/21.

(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) :** *INSTITUT MINES TELECOM Etablissement public* — FR et *BLOUET RAPHAEL* — FR.

(72) **Inventeur(s) :** ESSID SLIM et BLOUET RAPHAEL.

(73) **Titulaire(s) :** INSTITUT MINES TELECOM Etablissement public, BLOUET RAPHAEL.

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

(54) DISPOSITIF A CASQUE AUDIO PERFECTIONNE.

(57) L'invention concerne un traitement de données pour une restitution sonore sur un dispositif de restitution sonore (DIS), de type casque ou oreillettes, portable par un utilisateur dans un environnement (ENV). Le dispositif comporte au moins un haut-parleur (HP), au moins un microphone (MIC), et une connexion à un circuit de traitement comportant:

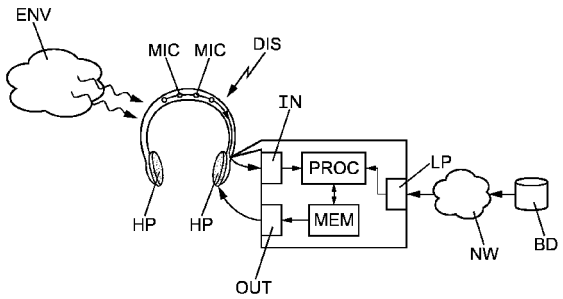
- une interface d'entrée (IN) pour recevoir des signaux issus au moins du microphone,
- une unité de traitement (PROC, MEM) pour lire au moins un contenu audio à restituer sur le haut-parleur, et
- une interface de sortie (OUT) pour délivrer au moins des signaux audio à restituer par le haut-parleur.

L'unité de traitement est agencée pour:

a) analyser les signaux issus du microphone pour identifier des sons émis par l'environnement et correspondant à des classes de sons cibles, prédéterminées,

b) sélectionner au moins un son identifié, selon un critère de préférence d'utilisateur, et

c) construire lesdits signaux audio à restituer par le haut-parleur, par un mixage choisi entre le contenu audio et le son sélectionné.



FR 3 059 191 - A1

Dispositif à casque audio, perfectionné

L'invention est relative à un dispositif portable d'écoute sonore. Il peut s'agir d'un casque audio avec des écouteurs gauche et droit, ou encore d'oreillettes portatives gauche et droite.

On connait des casques d'écoute audio antibruit, basés sur une captation par un réseau de microphones de l'environnement sonore de l'utilisateur. De manière générale, ces dispositifs cherchent à construire, en temps réel, le filtre optimal permettant de réduire au maximum la contribution de l'environnement sonore dans le signal sonore perçu par l'utilisateur. Il a été proposé récemment un filtre du bruit environnant qui peut être fonction du type d'environnement renseigné par l'utilisateur lui-même, lequel peut alors sélectionner différents modes d'annulation du bruit (bureau, extérieur, etc.). Le mode « extérieur » dans ce cas prévoit une réinjection du signal environnant (mais à un niveau beaucoup plus faible que sans filtre, et ce de manière à permettre à l'utilisateur de rester conscient de son environnement).

On connait aussi des casques audio et oreillettes, sélectifs, permettant une écoute personnalisée de l'environnement. Apparus récemment, ces produits permettent de modifier la perception de l'environnement selon deux axes :

- l'augmentation de la perception (intelligibilité de la parole), et
- la protection de l'appareil auditif en environnement bruité.

Il peut s'agir d'écouteurs audio, paramétrables via une application sur smartphone. L'amplification de la parole est possible dans un environnement bruité, la parole étant généralement localisée devant l'utilisateur.

Il peut s'agir aussi d'écouteurs audio connectés à un smartphone, permettant à l'utilisateur de configurer sa perception de l'environnement sonore : ajuster le volume, ajouter un équaliseur ou des effets sonores.

On peut citer aussi les casques et écouteurs interactifs, pour de la réalité augmentée, permettant d'enrichir l'environnement sonore (jeu, reconstitution historique) ou d'accompagner une activité de l'utilisateur (coach virtuel).

Enfin, les procédés mis en œuvre par certaines prothèses auditives pour améliorer l'expérience de l'utilisateur mal entendant proposent des axes d'innovation tels que l'amélioration de la sélectivité spatiale (suivant la direction des yeux de l'utilisateur par exemple).

Toutefois, ces différentes réalisations existantes ne permettent pas :

- d'analyser et interpréter l'activité de l'utilisateur, ni les contenus qu'il consomme, ni l'environnement (notamment la scène sonore) dans lequel il est immergé ;
- de modifier automatiquement le rendu audio en fonction du résultat de ces analyses.

Typiquement, les casques anti-bruit sont basés sur une captation multicanal exclusivement sonore de l'environnement de l'utilisateur. Ils cherchent à réduire de manière globale sa contribution dans le signal perçu par l'utilisateur indépendamment de la nature de l'environnement, et ce même s'il contient des informations potentiellement intéressantes. Ces dispositifs tendent donc à isoler l'utilisateur de son environnement.

Les prototypes de casque audio sélectifs permettent à l'utilisateur de configurer son environnement sonore par exemple en appliquant des filtres d'égalisation ou en augmentant l'intelligibilité de la parole. Ces dispositifs permettent d'améliorer la perception de l'environnement de l'utilisateur mais ne permettent pas de modifier réellement les contenus diffusés en fonction de l'état de l'utilisateur ou des classes de sons présentes dans l'environnement. Dans cette configuration, l'utilisateur écoutant de la musique avec un fort volume est toujours isolé de son environnement et le besoin

d'un dispositif permettant à l'utilisateur de capter les informations pertinentes dans son environnement est toujours présent.

Certes, les casques et écouteurs interactifs peuvent être équipés de capteurs permettant de charger et de diffuser des contenus associés à un lieu (dans le cadre d'une visite touristique par exemple) ou à une activité (jeu, entrainement sportif). Si certains appareils disposent même de capteurs inertiels ou physiologiques pour surveiller l'activité de l'utilisateur et si la diffusion de certains contenus peut dépendre alors des résultats de l'analyse des signaux issus de ces capteurs, les contenus diffusés ne résultent pas d'un processus de génération automatique prenant en compte l'analyse de la scène sonore environnant l'utilisateur et ne permettent pas de sélectionner automatiquement les composantes de cet environnement pertinentes pour l'utilisateur. Par ailleurs, les modes de fonctionnement sont statiques, et ne suivent pas automatiquement l'évolution au cours du temps de l'environnement sonore, et encore moins d'autres paramètres évolutifs tels qu'un état physiologique par exemple de l'utilisateur.

La présente invention vient améliorer la situation.

Elle propose à cet effet un procédé mis en œuvre par des moyens informatiques, de traitement de données pour une restitution sonore sur un dispositif de restitution sonore, de type casque ou oreillettes, portable par un utilisateur dans un environnement, le dispositif comportant :

- au moins un haut-parleur,
- au moins un microphone,
- une connexion à un circuit de traitement,

le circuit de traitement comportant :

- une interface d'entrée pour recevoir des signaux issus au moins du microphone,
- une unité de traitement pour lire au moins un contenu audio à restituer sur le haut-parleur, et

- une interface de sortie pour délivrer au moins des signaux audio à restituer par le haut-parleur.

En particulier, l'unité de traitement est agencée en outre pour mettre en œuvre les étapes :

a) analyser les signaux issus du microphone pour identifier des sons émis par l'environnement et correspondant à des classes de sons cibles, prédéterminées,

b) sélectionner au moins un son identifié, selon un critère de préférence d'utilisateur, et

c) construire lesdits signaux audio à restituer par le haut-parleur, par un mixage choisi entre le contenu audio et le son sélectionné.

Dans une forme de réalisation possible, le dispositif comporte une pluralité de microphones et l'analyse des signaux issus des microphones comporte en outre un traitement de séparation de sources sonores dans l'environnement appliqué aux signaux issus des microphones.

Par exemple, à l'étape c), le son sélectionné peut être :

- analysé au moins en fréquence et durée,
- rehaussé par filtrage après le traitement de séparation de sources, et mixé au contenu audio.

Dans une réalisation où le dispositif comporte au moins deux haut-parleurs et la restitution des signaux sur les haut-parleurs applique un effet sonore 3D, une position de source sonore, détectée dans l'environnement et émettant un son sélectionné, peut être prise en compte pour appliquer un effet de spatialisation sonore de la source dans le mixage.

Dans une réalisation, le dispositif peut comporter en outre une connexion à une interface homme machine à disposition d'un utilisateur pour entrer des préférences de sélection de sons de l'environnement (au sens général, comme on le verra plus loin) et le critère de préférence d'utilisateur est alors déterminé par apprentissage d'un historique des préférences entrées par l'utilisateur et stockées en mémoire.

Dans une réalisation (alternative ou complémentaire), le dispositif peut comporter en outre une connexion à une base de données de préférences d'utilisateurs et le critère de préférence d'utilisateur est déterminé alors par analyse du contenu de ladite base de données.

Le dispositif peut comporter en outre une connexion à un ou plusieurs capteurs d'état d'un utilisateur du dispositif, de sorte que le critère de préférence d'utilisateur tient compte d'un état courant de l'utilisateur, contribuant alors à une définition de « l'environnement » de l'utilisateur, au sens général.

Dans une telle réalisation, le dispositif peut comporter une connexion à un terminal mobile à disposition de l'utilisateur du dispositif, ce terminal comportant avantageusement un ou plusieurs capteurs d'état de l'utilisateur.

L'unité de traitement peut être agencée en outre pour sélectionner un contenu à lire parmi une pluralité de contenus, en fonction de l'état capté de l'utilisateur.

Dans une forme de réalisation, les classes de sons cibles, prédéterminées, peuvent comporter au moins des sons de paroles, dont les empreintes vocales sont préenregistrées.

En outre, à titre d'exemple, l'étape a) peut comporter optionnellement au moins l'une des opérations suivantes :

- construction et application d'un filtre dynamique pour une annulation du bruit dans les signaux issus du microphone ;
- localisation et isolation de sources sonores de l'environnement par application d'un traitement de séparation de sources appliqué à des signaux issus de plusieurs microphones, et exploitant par exemple une formation de voies (dite « beamforming »), pour identifier des sources d'intérêt (pour l'utilisateur du dispositif)

- extraire des paramètres propres à ces sources d'intérêt en vue d'une restitution ultérieure des sons captés et issus de ces sources d'intérêt dans un mixage audio spatialisé ;
- identification des différentes classes de son correspondant aux sources (dans différentes directions spatiales) par un système de classification (par exemple par réseaux de neurones profonds) de classes de son connues (parole, musique, bruit, etc.),
- et possible identification par d'autres techniques de classification de la scène sonore (par exemple, reconnaissance sonore d'un bureau, d'une rue en extérieur, de transports, etc.).

En outre, à titre d'exemple, l'étape c) peut comporter optionnellement au moins l'une des opérations suivantes :

- filtrage temporel, filtrage spectral et/ou filtrage spatial (par exemple filtrage de Wiener, et/ou algorithme Duet), pour rehausser, à partir d'un ou plusieurs flux audio captés par une pluralité de microphones, une source sonore donnée (en se basant sur les paramètres extrait par le module de séparation de sources précité) ;
- rendu audio 3D, par exemple à l'aide de techniques de filtrage HRTF (Head Related Transfer Functions).

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.

L'invention vise aussi un dispositif de restitution sonore, de type casque ou oreillettes, portable par un utilisateur dans un environnement, le dispositif comportant :

- au moins un haut-parleur,
- au moins un microphone,
- une connexion à un circuit de traitement,

le circuit de traitement comportant :

- une interface d'entrée pour recevoir des signaux issus au moins du microphone,

- une unité de traitement pour lire au moins un contenu audio à restituer sur le haut-parleur, et
- une interface de sortie pour délivrer au moins des signaux audio à restituer par le haut-parleur.

L'unité de traitement est agencée en outre pour :
- analyser les signaux issus du microphone pour identifier des sons émis par l'environnement et correspondant à des classes de sons cibles, prédéterminées,
- sélectionner au moins un son identifié, selon un critère de préférence d'utilisateur, et
- construire lesdits signaux audio à restituer par le haut-parleur, par un mixage choisi entre le contenu audio et le son sélectionné.

L'invention propose ainsi un système incluant un dispositif audio intelligent, intégrant par exemple un réseau de capteurs, au moins un haut-parleur et un terminal (e.g. *smartphone*). L'originalité de ce système est d'être capable de générer automatiquement, en temps réel, la "*bande son optimale*" de l'utilisateur, c'est-à-dire le contenu multimédia le mieux adapté à son environnement et à son état propre.

L'état propre d'un utilisateur peut être défini par :

i) un ensemble de préférences (type de musique, classes de son d'intérêt, etc.) ;

ii) son activité (au repos, au bureau, en entrainement sportif, etc.) ;

iii) ses états physiologiques (stress, fatigue, effort, etc.) et/ou socio-émotionnels (personnalité, humeur, émotions, etc.).

Le contenu multimédia généré peut comporter un contenu audio principal (à diffuser dans le casque) et éventuellement des contenus multimédias secondaires (textes, images, vidéo) qui peuvent être diffusés via le terminal de type smartphone.

Les différents éléments de contenu regroupent à la fois les éléments de la base de contenus de l'utilisateur (musiques, vidéo, etc., hébergées sur le terminal ou dans le *cloud*), le résultat de captations effectuées par un réseau de capteurs que comporte le

système et des éléments synthétiques générés par le système (notifications, « jingles » sonores ou textuels, bruit de confort, etc.).

Ainsi, le système peut analyser automatiquement l'environnement de l'utilisateur et prédire les composantes potentiellement d'intérêt pour l'utilisateur afin de les restituer de manière augmentée et contrôlée, en les superposant de façon optimale aux contenus consommés par celui-ci (typiquement la musique qu'il écoute).

La restitution effective des contenus prend en compte la nature des contenus et des composantes extraites de l'environnement (ainsi que l'état propre de l'utilisateur dans une forme de réalisation plus sophistiquée). Le flux sonore diffusé dans le casque n'est plus issu de deux sources concurrentes :

- une source principale (musique ou émission de radio ou autre), et

- une source perturbatrice (le bruit ambiant),

mais d'un ensemble de flux d'informations dont les contributions relatives sont ajustées en fonction de leur pertinence.

Ainsi, un message diffusé dans l'enceinte d'une gare sera restitué de manière à ce qu'il soit bien perçu par l'utilisateur alors même que celui-ci écoute de la musique à un niveau élevé, tout en réduisant le bruit ambiant non pertinent pour l'utilisateur. Cette possibilité est offerte par l'ajout d'un module de traitement intelligent intégrant notamment des algorithmes de séparation de sources et de classification de scènes sonores. L'avantage applicatif direct est d'une part de reconnecter l'utilisateur avec son environnement ou de l'avertir si une classe de sons ciblés est détectée, et d'autre part de générer automatiquement un contenu adapté à chaque instant aux attentes de l'utilisateur grâce à un moteur de recommandation prenant en charge les différents éléments de contenu, précités.

Il convient de rappeler que les dispositifs de l'état de l'art ne permettent pas d'identifier automatiquement chaque classe de son présente dans l'environnement de l'utilisateur pour associer à chacune d'elle un traitement conforme aux attentes de

l'utilisateur (par exemple une mise en avant d'un son, ou au contraire une réduction, la génération d'une alerte), en fonction de son identification dans l'environnement. L'état de l'art n'utilise pas d'analyse de la scène sonore, ni l'état de l'utilisateur ou son activité pour calculer le rendu sonore.

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

- la figure 1 illustre un dispositif selon l'invention, dans une première forme de réalisation,
- la figure 2 illustre un dispositif selon l'invention, dans une deuxième forme de réalisation, ici connecté à un terminal mobile,
- la figure 3 illustre les étapes d'un procédé selon une forme de réalisation de l'invention, et
- la figure 4 précise des étapes du procédé de la figure 3, selon une forme de réalisation particulière.

En référence à la figure 1, un dispositif DIS de restitution sonore (de type casque ou oreillettes), porté par exemple par un utilisateur dans un environnement ENV, comporte au moins :

- un (ou deux, dans l'exemple représenté) haut-parleurs HP,
- au moins un capteur, par exemple un microphone MIC (ou une rangée de microphones dans l'exemple représenté pour capter une directivité des sons issus de l'environnement), et
- une connexion à un circuit de traitement.

Le circuit de traitement peut être intégré directement au casque et être logé dans une enceinte d'un haut-parleur (comme illustré sur la figure 1), ou peut, dans la variante illustrée sur la figure 2, être implémenté dans un terminal TER de l'utilisateur, par

exemple un terminal mobile de type smartphone, ou encore être distribué entre plusieurs terminaux de l'utilisateur (un smartphone, et un objet connecté comportant éventuellement d'autres capteurs). Dans cette variante, la connexion entre le casque (ou les oreillettes) et le circuit de traitement dédié du terminal s'effectue par une connexion USB ou radiofréquence courte portée (par exemple par Bluetooth ou autre) et le casque (ou les oreillettes) est équipé d'un émetteur/récepteur BT1, communiquant avec un émetteur/récepteur BT2 que comporte le terminal TER. Une solution hybride dans laquelle le circuit de traitement est distribué entre l'enceinte du casque et un terminal est également possible.

Dans l'un ou l'autre des modes de réalisation ci-avant, le circuit de traitement comporte :

- une interface d'entrée IN, pour recevoir des signaux issus au moins du microphone MIC,
- une unité de traitement comportant typiquement un processeur PROC et une mémoire MEM, pour interpréter, relativement à l'environnement ENV, les signaux issus du microphone par apprentissage (par exemple par classification, ou encore par « matching » de type « finger printing » par exemple),
- une interface de sortie OUT pour délivrer au moins des signaux audio fonctions de l'environnement et à restituer par le haut-parleur.

La mémoire MEM peut stocker des instructions d'un programme informatique au sens de la présente invention, et éventuellement des données temporaires (de calcul ou autre), ainsi que des données durables, comme par exemple les préférences de l'utilisateur, ou encore des données de définition de modèles ou autres, comme on le verra plus loin.

L'interface d'entrée IN est, dans une forme de réalisation sophistiquée, reliée à un réseau de microphones, ainsi qu'à un capteur inertiel (prévu sur le casque ou dans le terminal) et la définition des préférences de l'utilisateur.

Les données de préférences de l'utilisateur peuvent être stockées localement dans la mémoire MEM, comme indiqué ci-dessus. En variante, elles peuvent être stockées, avec éventuellement d'autres données, dans une base de données distante DB accessible par une communication via un réseau local ou étendu NW. Un module de communication LP avec un tel réseau convenant pour cet effet peut être prévu dans le casque ou dans le terminal TER.

Avantageusement, une interface homme/machine peut permettre à l'utilisateur de définir et de mettre à jour ses préférences. Dans la réalisation de la figure 2 où le dispositif DIS est appairé avec le terminal TER, l'interface homme/machine peut simplement correspondre à un écran tactile du smartphone TER par exemple. Sinon, il peut être prévu une telle interface directement sur le casque.

Dans la réalisation de la figure 2 toutefois, il est avantageusement possible de tirer profit de la présence de capteurs supplémentaires dans le terminal TER pour enrichir la définition de l'environnement de l'utilisateur, au sens général. Ces capteurs supplémentaires peuvent être des capteurs physiologiques propres à l'utilisateur (mesure d'électroencéphalogramme, mesure du rythme cardiaque, podomètre, etc.) ou tous autres capteurs permettant d'améliorer la connaissance du couple environnement/état courant de l'utilisateur. De plus, cette définition peut inclure directement la notification par l'utilisateur lui-même de son activité, de son état propre et de son environnement.

La définition de l'environnement peut prendre en compte en outre :

- l'ensemble des contenus accessibles et un historique des contenus consultés (musiques, vidéos, émissions radios, etc.),
- des métadonnées (par exemple le genre, les occurrences d'écoute par morceau) associées à la librairie musicale de l'utilisateur peuvent aussi être associées ;

- par ailleurs, l'historique de navigation et des applications de son smartphone ;
- l'historique de sa consommation de contenus en streaming (via un fournisseur de service) ou en local ;
- les préférences et l'activité en cours de ses connexions sur les réseaux sociaux.

Ainsi, l'interface d'entrée peut, au sens général, être connectée à un ensemble de capteurs, et comprendre aussi des modules de connexion (notamment l'interface LP) pour la caractérisation de l'environnement de l'utilisateur, mais aussi de ses habitudes et préférences (historiques de consommations de contenus, activités en streaming et/ou réseaux sociaux).

On décrit ci-après en référence à la figure 3 le traitement qu'opère l'unité de traitement précitée, surveillant l'environnement et éventuellement l'état de l'utilisateur pour caractériser les informations pertinentes et susceptibles d'être restituées dans le flux multimédia de sortie. Dans une forme de réalisation, cette surveillance est mise en œuvre par l'extraction automatique, via des modules de traitement du signal et d'intelligence artificielle, notamment de *machine learning* (représentés par l'étape S7 dans la figure 3), de paramètres importants pour la création du flux multimédia de sortie. Ces paramètres, notés P1, P2,…, dans les figures peuvent être typiquement des paramètres d'environnement qui doivent être pris en compte pour la restitution sur haut-parleurs. Par exemple, si un son capté dans l'environnement est identifié comme étant un signal de parole à restituer :

- un premier ensemble de paramètres peut être des coefficients d'un filtre optimal (type filtre de Wiener) permettant de rehausser le signal de parole pour en augmenter l'intelligibilité ;

- un deuxième paramètre est la directivité du son capté dans l'environnement et à restituer par exemple à l'aide d'un rendu binaural (technique de restitution utilisant des fonctions de transfert de type HRTF) ;

- etc.

On comprendra ainsi que ces paramètres P1, P2, …, sont à interpréter comme des « descripteurs » de l'environnement et de l'état propre de l'utilisateur au sens général, qui alimentent un programme de génération de la « bande son optimale » pour cet utilisateur. Cette bande son est obtenue par composition de ses contenus, d'éléments de l'environnement et d'éléments synthétiques.

Au cours de la première étape S1, l'unité de traitement sollicite l'interface d'entrée pour collecter les signaux issus du microphone ou du réseau de microphones MIC que porte le dispositif DIS. Bien entendu, d'autres capteurs (d'inertie, ou autres) dans le terminal TER à l'étape S2, ou ailleurs à l'étape S3 (capteurs connectés de rythme cardiaque, EEG, etc.), peuvent communiquer leurs signaux à l'unité de traitement. Par ailleurs, des données d'informations autres que des signaux captés (préférences de l'utilisateur à l'étape S5, et/ou l'historique de consommation des contenus et des connexions aux réseaux sociaux à l'étape S6) peuvent être transmises par la mémoire MEM et/ou par la base de données BD à l'unité de traitement.

A l'étape S4, toutes ces données et signaux propres à l'environnement et l'état de l'utilisateur (appelés ci-après de façon générique « environnement ») sont collectés et interprétés par la mise en œuvre, à l'étape S7, d'un module informatique de décodage de l'environnement par intelligence artificielle. À cet effet, ce module de décodage peut utiliser une base d'apprentissage qui peut, par exemple, être distante et sollicitée à l'étape S8 via le réseau NW (et l'interface de communication LP), afin d'extraire des paramètres pertinents P1, P2, P3, …, à l'étape S9 qui modélisent l'environnement de manière générale.

Comme détaillé plus loin en référence à la figure 4, à partir de ces paramètres notamment, la scène sonore à restituer est générée à l'étape S10 et transmise sous la forme de signaux audio aux haut-parleurs HP à l'étape S11. Cette scène sonore peut être accompagnée éventuellement d'informations graphiques, par exemple des métadonnées, à afficher sur l'écran du terminal TER à l'étape S12.

Ainsi, il est procédé à une analyse des signaux d'environnement, avec :

- une identification de l'environnement en vue d'estimer des modèles de prédiction permettant de caractériser l'environnement de l'utilisateur et son état propre (ces modèles étant utilisés avec un moteur de recommandation comme on le verra plus loin en référence à la figure 4), et
- une analyse acoustique fine permettant de générer des paramètres plus précis et servant à la manipulation du contenu audio à restituer (séparation/rehaussement de sources sonores particulières, effets sonores, mixage, spatialisation, ou autres).

L'identification de l'environnement permet de caractériser, par apprentissage automatique, le couple environnement/état propre de l'utilisateur. Il s'agit principalement :

- de détecter si certaines classes de sons cibles, parmi plusieurs classes préenregistrées, sont présentes dans l'environnement de l'utilisateur et de déterminer, le cas échéant, leur direction de provenance. Initialement, les classes de son cibles peuvent être définies, une à une, par l'utilisateur via son terminal ou en utilisant des modes de fonctionnement prédéfinis ;
- de déterminer l'activité de l'utilisateur : repos, au bureau, en activité dans une salle de sport, ou autres ;
- de déterminer l'état émotionnel et physiologique de l'utilisateur (par exemple « en forme », d'après un podomètre, ou « stressé » d'après son EEG) ;
- de décrire les contenus qu'il consomme au moyen de techniques d'analyse par le contenu (techniques d'audition et de vision par ordinateur, et de traitement des langues naturelles).

L'analyse acoustique fine permet de calculer les paramètres acoustiques qui sont utilisés pour la restitution audio (par exemple en restitution 3D).

En référence maintenant à la figure 4, à l'étape S17, un moteur de recommandation est utilisé pour recevoir les descripteurs de « l'environnement », en particulier les classes d'événements sonores identifiés (paramètres P1, P2, ..), et fournir sur cette base un modèle de recommandation (ou une combinaison de modèles) à l'étape S19. À cet

effet, le moteur de recommandation peut utiliser la caractérisation des contenus de l'utilisateur et leur similarité à des contenus externes ainsi que des préférences de l'utilisateur, qui ont été enregistrées dans une base d'apprentissage à l'étape S15, et/ou des préférences standards d'autres utilisateurs à l'étape S18. L'utilisateur peut aussi intervenir à cette étape avec son terminal pour entrer une préférence à l'étape S24, par exemple par rapport à un contenu ou une liste de contenus à jouer.

À partir de l'ensemble de ces recommandations, il est choisi un modèle de recommandation pertinent en fonction de l'environnement et de l'état de l'utilisateur (par exemple dans le groupe des musiques rythmées, en situation de mouvement de l'utilisateur apparemment dans une salle de sport). Il est mis en œuvre ensuite un moteur de composition à l'étape S20, lequel combine les paramètres P1, P2…, au modèle de recommandation, pour élaborer un programme de composition à l'étape S21. Il s'agit ici d'une routine qui suggère par exemple :

- un type de contenu spécifique à rechercher dans les contenus de l'utilisateur,

- en tenant compte de son état propre (par exemple son activité) et de certains types de sons de l'environnement extérieur identifiés dans les paramètres P1, P2, …,

- à mixer au contenu, selon un niveau sonore et un rendu spatial (audio 3D) qui a été défini par le moteur de composition.

Le moteur de synthèse, à proprement parler, du signal sonore intervient à l'étape S22, pour élaborer les signaux à restituer aux étapes S11 et S12, à partir :

- des contenus de l'utilisateur (issus de l'étape S25 (en tant que sous étape de l'étape S6), bien entendu, l'un des contenus ayant été sélectionné à l'étape S21 par le moteur de composition,

- des signaux sonores captés dans l'environnement (S1, éventuellement de paramètres P1, P2, … dans le cas d'une synthèse des sons de l'environnement à restituer), et

- d'autres sons, possiblement synthétiques, de notifications (bip, cloche, ou autre), pouvant annoncer un évènement extérieur et à mixer au contenu à restituer (sélectionné à l'étape S21 à partir de l'étape S16),

avec éventuellement un rendu 3D défini à l'étape S23.

Ainsi, le flux généré est adapté aux attentes de l'utilisateur et optimisé en fonction du contexte de sa diffusion, selon trois étapes principales dans une forme de réalisation particulière :

- l'utilisation d'un moteur de recommandation pour filtrer et sélectionner en temps réel les éléments de contenu à mixer pour la restitution sonore (et possiblement visuelle aussi) d'un flux multimédia (dit de « réalité contrôlée ») ;
- l'utilisation d'un moteur de composition de média qui programme l'agencement temporel, fréquentiel et spatial des éléments de contenu, avec des niveaux sonores respectifs définis également ;
- l'utilisation d'un moteur de synthèse générant les signaux du rendu sonore (et éventuellement visuel), avec possiblement une spatialisation sonore, suivant le programme établi par le moteur de composition.

Le flux multimédia généré comporte au moins des signaux audio mais potentiellement des notifications textuelles, haptiques et ou visuelles. Les signaux audio comprennent un mixage :

- d'un contenu sélectionné dans la base de contenus de l'utilisateur (musiques, vidéo, etc.), entré comme préférence par l'utilisateur à l'étape S24, ou recommandé directement par le moteur de recommandation en fonction de l'état de l'utilisateur et de l'environnement,

avec éventuellement

- des sons captés par le réseau de capteurs MIC, sélectionnés dans l'environnement sonore (donc filtrés), rehaussés (par exemple par des techniques de séparation de source) et traités pour qu'ils soient de texture fréquentielle, d'intensité et de spatialisation, ajustées pour être injectés dans le mixage de façon opportune,

et

- des éléments synthétiques récupérés d'une base à l'étape S16, par exemple des sons de notifications/jingles sonores/textuels, bruit de confort, etc.).

Le moteur de recommandation se base conjointement sur :

- les préférences de l'utilisateur obtenues de manière explicite à travers une forme de questionnement, ou de manière implicite en exploitant le résultat du décodage de son état propre,
- des techniques de filtrage collaboratif et de graphes sociaux, exploitant les modèles de plusieurs utilisateurs à la fois (étape S18),
- la description des contenus de l'utilisateur et leur similarité, afin de construire des modèles permettant de décider quels éléments de contenu doivent être joués à l'utilisateur.

Les modèles sont mis à jour de façon continue au cours du temps pour s'adapter à l'évolution de l'utilisateur.

Le moteur de composition planifie :

- l'instant auquel doit être joué chaque élément de contenu, notamment l'ordre dans lequel les contenus de l'utilisateur sont présentés (par exemple, l'ordre des morceaux de musique dans une playlist), et les moments ou les sons extérieurs ou les notifications sont diffusées : en temps réel ou en différé (par exemple entre deux morceaux d'une playlist) pour ne pas perturber l'écoute ou l'activité en cours de l'utilisateur à un moment inopportun ;
- la position spatiale (en vue d'un rendu 3D) de chaque élément de contenu ;
- les différents effets audio (gain, filtrage, égalisation, compression dynamique, écho ou réverbération (« reverb »), ralentissement/accélération temporelle, transposition…) qui doivent être appliqués à chaque élément de contenu.

La planification se base sur des modèles et des règles construites à partir du décodage de l'environnement de l'utilisateur et de son état propre. Par exemple, la position spatiale d'un évènement sonore capturé par les micros et le niveau de gain qui lui est associé dépendent du résultat de la détection de localisation de sources sonores que réalise le décodage de l'environnement à l'étape S7 de la figure 3.

Le moteur de synthèse s'appuie sur des techniques de traitement du signal, des langues naturelles et des images, respectivement pour la synthèse de sorties audio, textuelles et visuelles (images ou vidéos), et conjointement pour la génération de sorties multimédia, par exemple vidéo.

Dans le cas de la synthèse de la sortie audio, des techniques de filtrage temporel, spectral et/ou spatial peuvent être exploitées. Par exemple, la synthèse est d'abord réalisée localement sur des fenêtres temporelles courtes et le signal est reconstruit par addition-recouvrement avant d'être transmis à au moins deux haut-parleurs (un pour chaque oreille). Des gains (niveaux de puissance) et des effets audio différents sont appliqués aux différents éléments de contenu, tel que prévu par le moteur de composition.

Dans une réalisation particulière, le traitement appliqué par fenêtres peut inclure un filtrage (par exemple de Wiener) permettant de rehausser, à partir d'un ou plusieurs des flux audio captés, une source sonore particulière (telle que prévue par le moteur de composition).

Dans une réalisation particulière, le traitement peut inclure un rendu audio 3D, éventuellement à l'aide de techniques de filtrage HRTF (fonctions de transfert HRTF pour « Head Related Transfer Functions »).

Dans un premier exemple illustrant une implémentation minimale,

- la description de l'environnement de l'utilisateur est limitée à son environnement sonore ;
- l'état propre de l'utilisateur est limité à ses préférences : classe de son cible, notifications qu'il souhaite recevoir, ces préférences étant définies par l'utilisateur à l'aide de son terminal ;

- le dispositif (éventuellement en coopération avec le terminal) est équipé de capteurs inertiels (accéléromètre, gyroscope et magnétomètre) ;
- les paramètres de restitution sont automatiquement modifiés lorsqu'une classe de sons cibles est détectée dans l'environnement de l'utilisateur ;
- les messages de courtes durées peuvent être enregistrés ;
- des notifications peuvent être envoyées à l'utilisateur pour l'avertir de la détection d'un événement d'intérêt.

Les signaux captés sont analysés afin de déterminer :

- les classes de sons présentes dans l'environnement de l'utilisateur et les directions d'où elles proviennent, avec, à cet effet :
  - une détection des directions de plus fortes énergies sonore en analysant les contenus dans chacune de ces directions de manière indépendante,
  - une détermination globale pour chaque direction de la contribution de chacune des classes de son (par exemple en utilisant une technique de séparation de sources),
- les paramètres de modèles décrivant l'environnement de l'utilisateur et ceux des paramètres alimentant le moteur de recommandation.

Dans un deuxième exemple illustrant une implémentation plus sophistiquée, un ensemble de capteurs comprenant un réseau de microphones, une caméra vidéo, un podomètre, des capteurs inertiels (accéléromètres, gyroscopes, magnétomètres), des capteurs physiologiques peuvent capter l'environnement visuel et sonore de l'utilisateur (micros et caméra), les données caractérisant son mouvement (capteurs inertiels, podomètre) et ses paramètres physiologiques (EEG, ECG, EMG, électrodermal) ainsi que l'ensemble des contenus qu'il est en train de consulter (musiques, émissions radio, vidéos, historique de navigation et des applications de son smartphone). Ensuite, les différents flux sont analysés pour extraire l'information liée à l'activité de l'utilisateur, son humeur, son état de fatigue et son environnement (par exemple course sur tapis

roulant dans une salle de sport, de bonne humeur et en état de faible fatigue). Un flux musical adapté à l'environnement et à l'état propre de l'utilisateur peut être généré (par exemple une playlist dont chaque morceau est sélectionné en fonction de ses goûts musicaux, de sa foulée et de son état de fatigue). Alors que toutes les sources sonores sont annulées dans le casque de l'utilisateur, la voix d'un entraineur (« coach sportif ») à proximité de l'utilisateur, lorsqu'elle est identifiée (empreinte vocale préalablement enregistrée), est mixée au flux et restituée spatialement à l'aide de techniques de rendu binaural (par HRTF par exemple).

REVENDICATIONS

1. Procédé mis en œuvre par des moyens informatiques, de traitement de données pour une restitution sonore sur un dispositif de restitution sonore, de type casque ou oreillettes, portable par un utilisateur dans un environnement, le dispositif comportant :
- au moins un haut-parleur,
- au moins un microphone,
- une connexion à un circuit de traitement,
le circuit de traitement comportant :
- une interface d'entrée pour recevoir des signaux issus au moins du microphone,
- une unité de traitement pour lire au moins un contenu audio à restituer sur le haut-parleur, et
- une interface de sortie pour délivrer au moins des signaux audio à restituer par le haut-parleur,
caractérisé en ce que l'unité de traitement est agencée en outre pour mettre en œuvre les étapes :
a) analyser les signaux issus du microphone pour identifier des sons émis par l'environnement et correspondant à des classes de sons cibles, prédéterminées,
b) sélectionner au moins un son identifié, selon un critère de préférence d'utilisateur, et
c) construire lesdits signaux audio à restituer par le haut-parleur, par un mixage choisi entre le contenu audio et le son sélectionné.

2. Procédé selon la revendication 1, caractérisé en ce que, le dispositif comportant une pluralité de microphones, l'analyse des signaux issus des microphones comporte en outre un traitement de séparation de sources sonores dans l'environnement appliqué aux signaux issus des microphones.

3. Procédé selon la revendication 2, caractérisé en ce que, à l'étape c), le son sélectionné est :
- analysé au moins en fréquence et durée,

- rehaussé par filtrage après le traitement de séparation de sources, et mixé au contenu audio.

4. Procédé selon l'une des revendications 2 et 3, caractérisé en ce que, le dispositif comportant au moins deux haut-parleurs et la restitution des signaux sur les haut-parleurs appliquant un effet sonore 3D, une position de source sonore, détectée dans l'environnement et émettant un son sélectionné, est prise en compte pour appliquer un effet de spatialisation sonore de la source dans le mixage.

5. Procédé selon l'une des revendications précédentes, caractérisé en ce que le dispositif comporte en outre une connexion à une interface homme machine à disposition d'un utilisateur pour entrer des préférences de sélection de sons de l'environnement, et en ce que le critère de préférence d'utilisateur est déterminé par apprentissage d'un historique des préférences entrées par l'utilisateur et stockées en mémoire.

6. Procédé selon l'une des revendications précédentes, caractérisé en ce que le dispositif comporte en outre une connexion à une base de données de préférences d'utilisateurs et le critère de préférence d'utilisateur est déterminé par analyse du contenu de ladite base de données.

7. Procédé selon l'une des revendications précédentes, caractérisé en ce que le dispositif comporte en outre une connexion à un ou plusieurs capteurs d'état d'un utilisateur du dispositif, et en ce que le critère de préférence d'utilisateur tient compte d'un état courant de l'utilisateur.

8. Procédé selon la revendication 7, caractérisé en ce que le dispositif comporte une connexion à un terminal mobile à disposition de l'utilisateur du dispositif, le terminal comportant un ou plusieurs capteurs d'état de l'utilisateur.

9. Procédé selon l'une des revendications 7 et 8, caractérisé en ce que, l'unité de traitement est agencée en outre pour sélectionner un contenu à lire parmi une pluralité de contenus, en fonction de l'état de l'utilisateur.

10. Procédé selon l'une des revendications précédentes, caractérisé en ce que les classes de sons cibles, prédéterminées, comportent au moins des sons de paroles, d'empreintes vocales préenregistrées.

11. Programme informatique caractérisé en ce qu'il comporte des instructions pour la mise en œuvre du procédé selon l'une des revendications 1 à 10, lorsque ce programme est exécuté par un processeur.

12. Dispositif de restitution sonore, de type casque ou oreillettes, portable par un utilisateur dans un environnement, le dispositif comportant :
- au moins un haut-parleur,
- au moins un microphone,
- une connexion à un circuit de traitement,
le circuit de traitement comportant :
- une interface d'entrée pour recevoir des signaux issus au moins du microphone,
- une unité de traitement pour lire au moins un contenu audio à restituer sur le haut-parleur, et
- une interface de sortie pour délivrer au moins des signaux audio à restituer par le haut-parleur,
caractérisé en ce que l'unité de traitement est agencée en outre pour :
- analyser les signaux issus du microphone pour identifier des sons émis par l'environnement et correspondant à des classes de sons cibles, prédéterminées,
- sélectionner au moins un son identifié, selon un critère de préférence d'utilisateur, et
- construire lesdits signaux audio à restituer par le haut-parleur, par un mixage choisi entre le contenu audio et le son sélectionné.

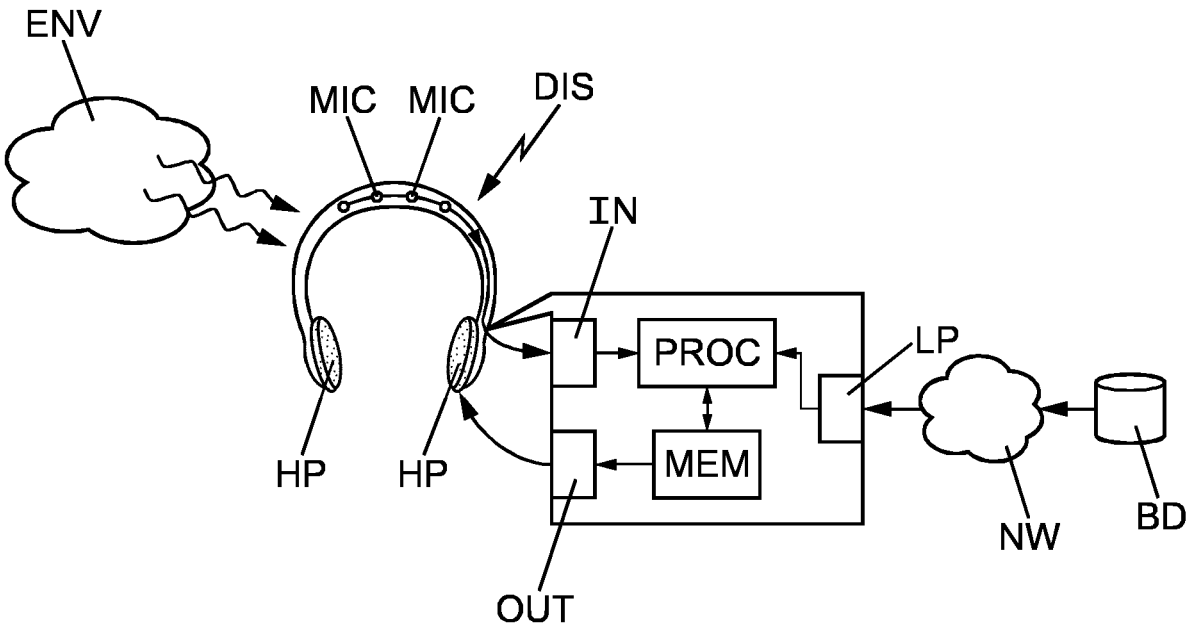

ENV
MIC
MIC
DIS
IN
PROC
LP
HP
HP
MEM
NW
BD
OUT

FIG. 1

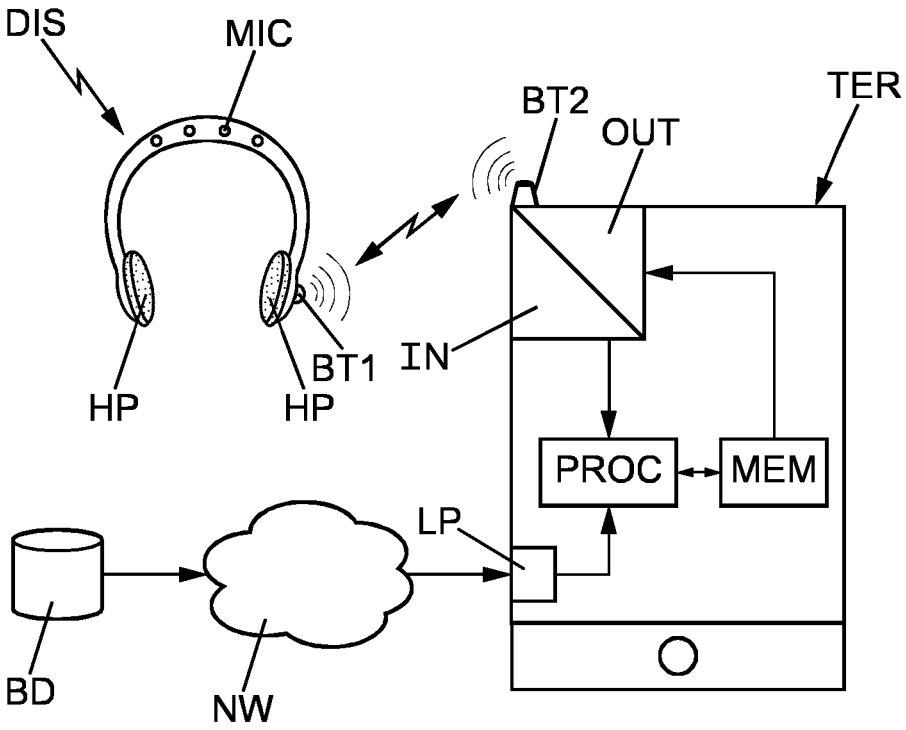

DIS
MIC
BT2
OUT
TER
HP
HP
BT1
IN
PROC
MEM
LP
BD
NW

FIG. 2

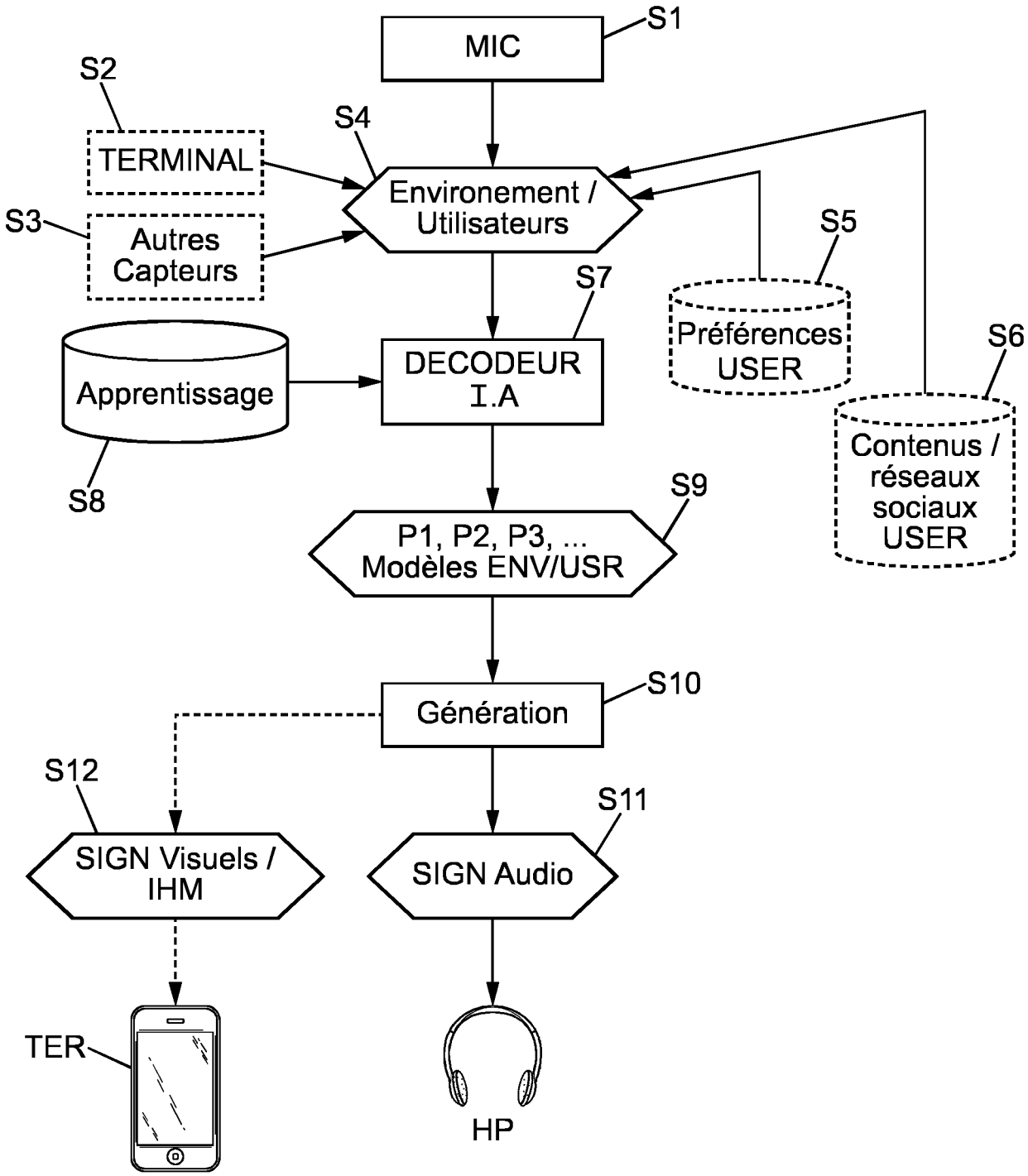

MIC
S1
S2
TERMINAL
S4
Environement / Utilisateurs
S3
Autres Capteurs
S5
S7
Préférences USER
S6
Apprentissage
DECODEUR I.A
Contenus / réseaux sociaux USER
S8
S9
P1, P2, P3, ... Modèles ENV/USR
Génération
S10
S12
S11
SIGN Visuels / IHM
SIGN Audio
TER
HP

FIG. 3

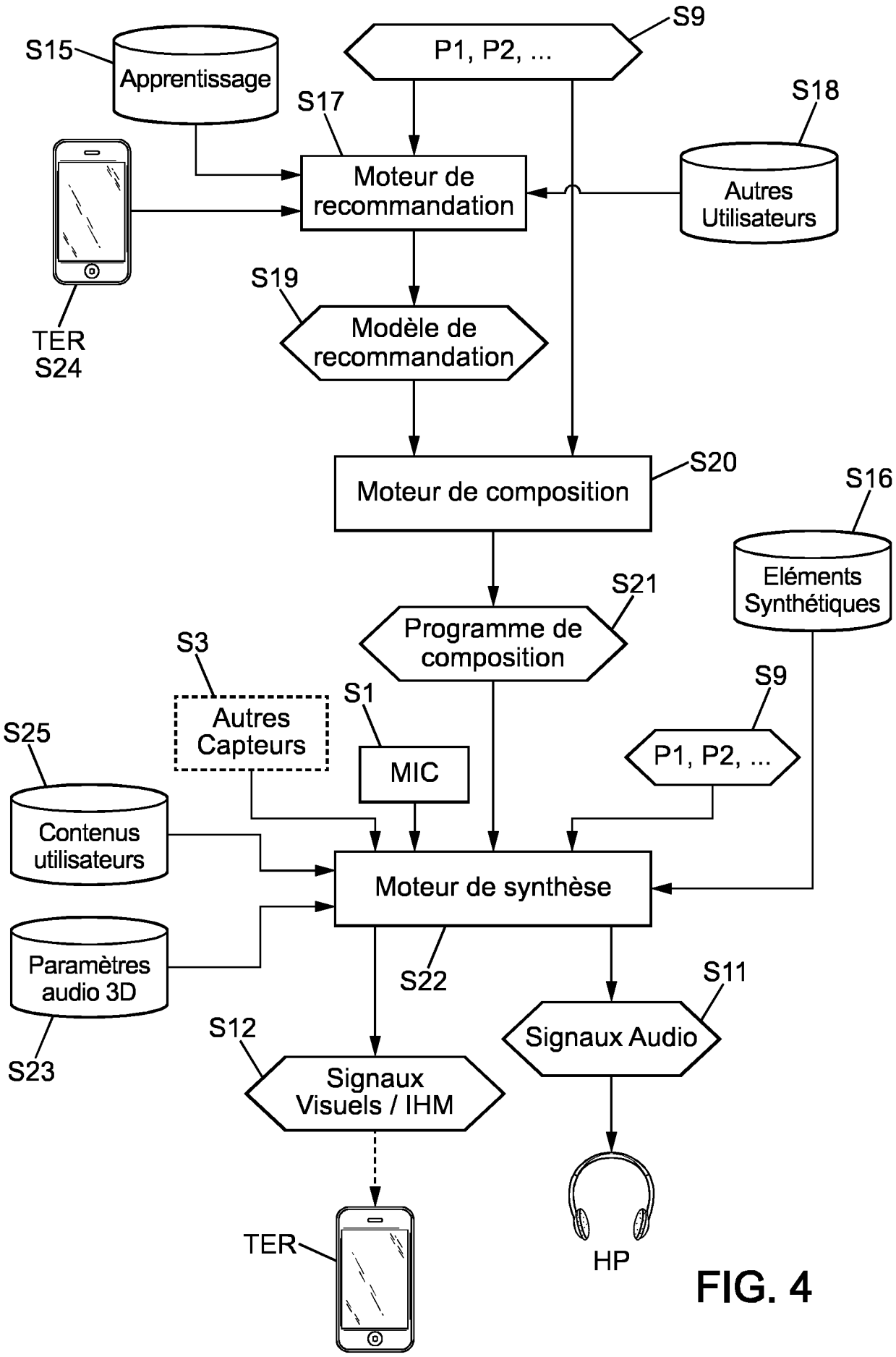

S15
Apprentissage
S9
P1, P2, ...
S17
S18
Moteur de recommandation
Autres Utilisateurs
TER
S24
S19
Modèle de recommandation
S20
Moteur de composition
S16
Eléments Synthétiques
S21
Programme de composition
S3
Autres Capteurs
S1
MIC
S9
P1, P2, ...
S25
Contenus utilisateurs
Moteur de synthèse
Paramètres audio 3D
S23
S22
S11
Signaux Audio
S12
Signaux Visuels / IHM
TER
HP

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 834006
FR 1661324

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	US 2016/163303 A1 (BENATTAR BENJAMIN D [US] ET AL) 9 juin 2016 (2016-06-09) * alinéas [0035] - [0046], [0118] - [0157]; figure 9 *	1-12	H04R5/033 H04R3/00 G10K11/16
A	EP 2 602 728 A1 (FRANCE TELECOM [FR]) 12 juin 2013 (2013-06-12) * revendications 1-3 *	5	
A	US 2015/222989 A1 (LABROSSE JEAN-PAUL [US] ET AL) 6 août 2015 (2015-08-06) * alinéa [0061]; revendications 1,2,9 *	7-9	
A	US 2015/078575 A1 (SELIG AARON ALEXANDER [US] ET AL) 19 mars 2015 (2015-03-19) * alinéas [0025] - [0043] *	1-12	

DOMAINES TECHNIQUES RECHERCHÉS (IPC)

H04R
H04S
G10K

Date d'achèvement de la recherche	Examinateur
26 juillet 2017	Van Hoorick, Jan

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 1661324 FA 834006

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 26-07-2017
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
US 2016163303 A1	09-06-2016	AUCUN	
EP 2602728 A1	12-06-2013	EP 2602728 A1	12-06-2013
		FR 2983605 A1	07-06-2013
US 2015222989 A1	06-08-2015	AUCUN	
US 2015078575 A1	19-03-2015	US 2015078575 A1	19-03-2015
		US 2016234589 A1	11-08-2016

EPO FORM P0465