(19) RÉPUBLIQUE FRANÇAISE

INSTITUT NATIONAL DE LA PROPRIÉTÉ INDUSTRIELLE

COURBEVOIE

(11) **N° de publication :** ***3 137 520***

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

(21) **N° d'enregistrement national :** **22 06706**

(51) Int Cl[8] : ***H 04 N 21/43*** *(2022.01), H 04 N 21/439*

(12) DEMANDE DE BREVET D'INVENTION A1

(22) **Date de dépôt :** 01.07.22.

(30) **Priorité :**

(43) **Date de mise à la disposition du public de la demande :** 05.01.24 Bulletin 24/01.

(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) :** *ORANGE Société anonyme* — FR.

(72) **Inventeur(s) :** DAUVIN Laurent, MEUNIER David, BERTHOUT Alban et BRECHET Matthieu.

(73) **Titulaire(s) :** ORANGE Société anonyme.

(74) **Mandataire(s) :** CABINET VIDON BREVETS & STRATÉGIE.

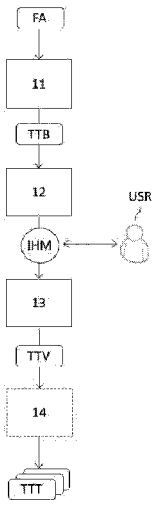
(54) Procédé de génération dynamique d'une transcription textuelle d'un flux audio diffusé en continu.

(57) Procédé de génération dynamique d'une transcription textuelle d'un flux audio diffusé en continu .

L'invention se rapporte à un procédé de génération dynamique d'une transcription textuelle d'un flux audio (FA) diffusé en continu. Ce procédé, mis en œuvre au niveau d'un dispositif recevant le flux audio, comprend:

- le traitement (11) du flux audio (FA) par un système de reconnaissance vocale, délivrant une première transcription textuelle dudit flux audio (TTB);
- la fourniture (12) de ladite première transcription textuelle (TTB) à un module de validation de transcription textuelle comprenant au moins une interface homme-machine (IHM) de validation de ladite première transcription textuelle;
- l'obtention (13), en provenance du module de validation, d'une deuxième transcription textuelle du flux audio, dite transcription textuelle validée (TTV).

Figure d'abrégé: Figure 1



FR 3 137 520 - A1

Description

Titre de l'invention : Procédé de génération dynamique d'une transcription textuelle d'un flux audio diffusé en continu.

Domaine technique

[0001] L'invention se situe dans le domaine de la transcription textuelle de contenus audio. Plus particulièrement, l'invention se rapporte à la transcription textuelle dynamique de contenus audio diffusés en direct.

Art antérieur

[0002] Avec le nombre sans cesse croissant de contenus audio et vidéo accessibles en direct ou en différé (par exemple via le réseau Internet), et la multiplication des équipements permettant de les consommer (télévisions, ordinateurs, smartphones, tablettes, etc.), y compris en situation de mobilité, les besoins en transcription textuelle de la parole se sont multipliés.

[0003] La fourniture de telles transcriptions, par exemple sous forme de sous-titres intégrés au sein d'un contenu vidéo, ou plus simplement d'une retranscription textuelle sur un écran d'un contenu audio, présente en effet un intérêt indéniable dans de nombreuses situations. Elle permet par exemple à un utilisateur de prendre connaissance de l'information associée à un contenu audio ou vidéo sur un terminal électronique, et ceci même avec le son coupé, ce qui peut être utile pour éviter de déranger les autres usagers d'un lieu public ou d'un transport en commun. Elle permet également de rendre certains contenus audios compréhensibles par un plus large public, par exemple en proposant un sous-titrage dans une langue différente de celle dans laquelle le contenu est diffusé. Dans certaines situations et/ou pour certains contenus, la présence d'une telle transcription peut même s'avérer obligatoire, dans le cadre par exemple du respect de lois pour l'égalité des droits et des chances pour tous, de manière à ce que des personnes sourdes ou malentendantes puissent avoir accès à l'information au même titre et dans les mêmes conditions que les personnes entendantes.

[0004] La génération de transcriptions textuelles de contenus préenregistrés, destinés à être diffusés en différé, ne pose généralement pas de problème particulier : le temps n'étant dans ce cas pas un facteur critique, la production d'une transcription de bonne qualité, et sa synchronisation, le cas échéant, avec un contenu vidéo associé, reposent sur des techniques maintenant bien développées, qui peuvent être mises en œuvre à des coûts maitrisés. Il en va cependant autrement lorsque la génération de telles transcriptions textuelles porte sur des contenus diffusés en direct (souvent également désignés sous le terme anglais de contenus « *live* »), puisque la transcription doit être générée rapidement, à la volée, au fur et à mesure de la diffusion du flux audio.

[0005] Pour parvenir à mettre en œuvre de telles transcriptions en temps réel (ou à tout le moins en quasi temps-réel), il existe actuellement deux solutions principales.

[0006] Une première solution consiste à utiliser un système de reconnaissance vocale (également souvent désigné sous le terme de ASR, de l'anglais « *Automatic Speech Recognition* ») pour obtenir et restituer de manière automatique une transcription textuelle de la parole. Une telle solution, qui repose par exemple sur l'utilisation de réseau de neurones artificiels, est encore peu fiable en ce qui concerne la qualité de la transcription obtenue, avec un taux d'erreur de mots (« *word error rate* » en anglais) qui est susceptible de rester relativement élevé dans certaines situations (par exemple en cas de diction ou d'accent particuliers du locuteur, de propos associés à un champs lexical complexe ou très spécifique, etc.). Ce manque de fiabilité est problématique, notamment quand il s'agit de retranscrire en temps-réel des allocutions diffusées en direct sur des sujets importants et potentiellement sensibles, comme peuvent l'être par exemple certaines interventions d'hommes d'état portant sur des risques sanitaires, environnementaux, économiques, politiques ou militaires, etc.

[0007] Une deuxième solution consiste à recourir aux services de personnes spécialisées dans la restitution en direct et à l'écrit de la parole, connues sous le nom de vélotypistes. Une telle solution permet d'obtenir une transcription textuelle de meilleure qualité et plus fiable que la première solution, en s'appuyant notamment sur les capacités humaines à contextualiser le propos, et éventuellement à le reformuler ou l'adapter pour plus de clarté à l'écrit. Cette solution est cependant très onéreuse, d'une part parce que les spécialistes vélotypistes sont rares, et d'autre part parce que chaque mission de transcription exige de leur part un important travail de préparation en amont consistant notamment à s'approprier le contexte, identifier le vocabulaire susceptible d'être utilisé par l'orateur, et configurer à l'avance les outils utilisés en conséquence, de manière à être prêt pour délivrer un service de qualité lors du direct.

[0008] Ces deux solutions existantes présentent donc l'une comme l'autre des inconvénients qui leur sont propres. Par ailleurs, lorsque qu'elles sont utilisées pour réaliser la transcription du contenu audio d'un flux vidéo diffusé en direct, l'affichage au fil de l'eau (par exemple sous la forme de sous-titres) de la transcription textuelle générée pâtit d'un léger décalage temporel vis-à-vis du contenu audio et vidéo, imputable au temps de traitement nécessaire à la machine (dans le cas de la première solution) ou au vélotypiste (dans le cas de la deuxième solution) pour générer cette transcription.

[0009] Il existe donc un besoin pour une solution permettant d'offrir un compromis plus satisfaisant entre qualité de transcription et coût de mise en œuvre pour la génération de transcriptions textuelles associées à des contenus diffusés en direct, c'est-à-dire à des contenus diffusés en continu au fur et à mesure qu'ils sont produits. Plus particulièrement, il existe un besoin pour une solution de transcription textuelle en direct qui

délivre des résultats de qualité suffisante pour répondre à la majorité des usages, y compris dans des contextes particulièrement exigeants, tout en restant à des coûts maitrisés.

Résumé de l'invention

[0010] La présente technique permet de proposer une solution visant à remédier à certains inconvénients de l'art antérieur. Selon un aspect, la présente technique se rapporte en effet à un procédé de génération dynamique d'une transcription textuelle d'un flux audio diffusé en continu. Un tel procédé comprend, au niveau d'un dispositif recevant ledit flux audio :

[0011] - le traitement dudit flux audio par un système de reconnaissance vocale, délivrant une première transcription textuelle dudit flux audio ;

[0012] - la fourniture de ladite première transcription textuelle à un module de validation de transcription textuelle comprenant au moins une interface homme-machine de validation de ladite première transcription textuelle ;

[0013] - l'obtention, en provenance dudit module de validation, d'une deuxième transcription textuelle dudit flux audio, dite transcription textuelle validée.

[0014] De cette manière, la présente technique offre une alternative intéressante aux solutions de l'art antérieur, en proposant une solution hybride permettant d'obtenir une transcription textuelle de meilleure qualité que celle qui aurait été obtenue via un système de reconnaissance vocale utilisé seul, et à des coûts moins élevés que celle qui aurait été obtenue via le recours à des services de vélotypie pure.

[0015] Dans un mode de réalisation particulier, le procédé de génération dynamique d'une transcription textuelle comprend en outre une traduction de ladite transcription textuelle validée, délivrant au moins une troisième transcription textuelle dudit flux audio dans une langue autre qu'une langue d'origine associée audit flux audio, dite transcription textuelle traduite.

[0016] De cette manière, la transcription textuelle obtenue avec la présente technique peut être disponible dans plusieurs langues, favorisant ainsi la compréhension par un plus large public du contenu diffusé en continu.

[0017] Dans un mode de réalisation particulier, ledit flux audio correspond à une piste audio d'un flux vidéo source diffusé en continu.

[0018] De cette manière, la présente technique est notamment adaptée à la génération à la volée de sous-titres associés à un contenu vidéo.

[0019] Selon une caractéristique particulière de ce mode de réalisation, le procédé de génération dynamique d'une transcription textuelle comprend en outre :

[0020] - la mise en mémoire tampon dudit flux vidéo source pendant une temporisation d'une durée prédéterminée, délivrant un flux vidéo temporisé ;

[0021] - l'intégration à la volée de ladite transcription textuelle validée et/ou de ladite au moins une transcription textuelle traduite au sein dudit flux vidéo temporisé, délivrant un flux vidéo temporisé enrichi.

[0022] De cette manière, en temporisant la diffusion du flux vidéo source d'un temps par exemple au moins égal au temps de traitement nécessaire à la génération d'une transcription textuelle d'au moins une partie du contenu diffusé, la présente technique permet de réduire ou d'éliminer le décalage temporel, généralement observé avec les solutions de l'art antérieur, entre le moment où un segment audio est diffusé et le moment où la transcription textuelle associée est affichée.

[0023] Dans un mode de réalisation particulier, ladite durée prédéterminée est inférieure à une minute.

[0024] De cette manière, la diffusion du flux vidéo reste effectuée dans des conditions de direct ou à tout le moins de « quasi-direct », la temporisation du flux vidéo étant de faible durée.

[0025] Dans un mode de réalisation particulier, ladite intégration de la transcription textuelle au sein du flux vidéo temporisé comprend l'ajout à la volée de métadonnées à ce flux, lesdites métadonnées comprenant ladite transcription textuelle validée et/ou ladite au moins une transcription textuelle traduite.

[0026] De cette manière, la présente technique offre beaucoup de souplesse au consommateur final du contenu, en termes de restitution de la transcription textuelle. Plus particulièrement, l'intégration de la transcription textuelle sous forme de métadonnées offre l'avantage, vis-à-vis d'autres solutions telles par exemple qu'une incrustation de texte directement dans le flux vidéo, de permettre au consommateur du contenu de choisir s'il souhaite ou non afficher ladite transcription, voire de choisir la langue d'affichage de ladite transcription le cas échéant.

[0027] Selon une caractéristique particulière de ce mode de réalisation, ledit ajout de métadonnées audit flux vidéo temporisé tient compte de données d'horodatage associées au traitement mis en œuvre par ledit système de reconnaissance vocale.

[0028] De cette manière, l'affichage de la transcription textuelle, par exemple sous forme de sous-titres, est parfaitement synchronisée avec la piste audio du flux vidéo diffusé.

[0029] Selon un autre aspect, la présente technique se rapporte également à un dispositif de génération dynamique d'une transcription textuelle d'un flux audio diffusé en continu. Un tel dispositif comprend :

[0030] - des moyens de traitement dudit flux audio, comprenant un système de reconnaissance vocale, délivrant une première transcription textuelle dudit flux audio ;

[0031] - des moyens de fourniture de ladite première transcription textuelle à un module de validation de transcription textuelle comprenant au moins une interface homme-machine de validation de ladite première transcription textuelle ;

[0032] - des moyens d'obtention, en provenance dudit module de validation, d'une deuxième transcription textuelle dudit flux audio, dite transcription textuelle validée.

[0033] Selon un autre aspect, la technique proposée se rapporte également à un produit programme d'ordinateur téléchargeable depuis un réseau de communication et/ou stocké sur un support lisible par ordinateur et/ou exécutable par un microprocesseur, comprenant des instructions de code de programme pour l'exécution d'un procédé de génération dynamique d'une transcription textuelle d'un flux audio diffusé en continu tel que décrit précédemment, lorsqu'il est exécuté sur un ordinateur.

[0034] La technique proposée vise également un support d'enregistrement lisible par un ordinateur sur lequel est enregistré un programme d'ordinateur comprenant des instructions de code de programme pour l'exécution des étapes du procédé tel que décrit précédemment, dans l'un quelconque de ses modes de réalisation.

[0035] Un tel support d'enregistrement peut être n'importe quelle entité ou dispositif capable de stocker le programme. Par exemple, le support peut comporter un moyen de stockage, tel qu'une ROM, par exemple un CD ROM ou une ROM de circuit microélectronique, ou encore un moyen d'enregistrement magnétique, par exemple une clé USB ou un disque dur.

[0036] D'autre part, un tel support d'enregistrement peut être un support transmissible tel qu'un signal électrique ou optique, qui peut être acheminé via un câble électrique ou optique, par radio ou par d'autres moyens, de sorte que le programme d'ordinateur qu'il contient est exécutable à distance. Le programme selon l'invention peut être en particulier téléchargé sur un réseau, par exemple le réseau Internet.

[0037] Les différents modes de réalisation mentionnés ci-dessus sont combinables entre eux pour la mise en œuvre de l'invention.

Figures

[0038] D'autres caractéristiques et avantages de l'invention apparaîtront plus clairement à la lecture de la description suivante d'un mode de réalisation préférentiel, donné à titre de simple exemple illustratif et non limitatif, et des dessins annexés, parmi lesquels :

[0039] [Fig.1] illustre le principe général d'un procédé de génération dynamique d'une transcription textuelle d'un flux audio diffusé en continu, dans un mode de réalisation particulier de la technique proposée ;

[0040] [Fig.2] présente de manière schématique les différents blocs fonctionnels d'un dispositif de génération dynamique d'une transcription textuelle d'un flux vidéo diffusé en continu, dans un mode de réalisation particulier de la technique proposée ;

[0041] [Fig.3] décrit de manière schématique un mécanisme de synchronisation d'une transcription textuelle avec un flux audio ou vidéo auquel elle est associée, dans un mode de réalisation particulier de la technique proposée ;

[0042] [Fig.4] présente une architecture simplifiée d'un dispositif de génération dynamique d'une transcription textuelle pour la mise en œuvre de la technique proposée, dans un mode de réalisation particulier.

Description détaillée de l'invention

[0043] La présente demande permet de remédier à certains des inconvénients précités des solutions de l'art antérieur pour la génération de transcriptions textuelles relatives à des contenus diffusés en direct.

[0044] La technique proposée vise notamment à proposer une solution offrant un compromis particulièrement intéressant entre qualité de la transcription textuelle obtenue et coût de mise en œuvre, que n'offrent pas les solutions de l'art antérieur.

[0045] La présente technique portant sur des flux diffusés en direct, on entend par traitement « par segment » (e.g. transmission par segment, délivrance par segment, etc.) dans la suite du présent document un traitement réalisé au fil de l'eau, au fur et à mesure de la réception d'un flux de données audio ou textuelles émis en continu, typiquement un traitement par mot ou par groupe de mots (par exemple par « groupe de souffle », i.e. par groupe de mots prononcés dans un seul souffle, sans pause) réalisé à la volée.

[0046] Par ailleurs, le terme « module » peut correspondre aussi bien à un composant logiciel qu'à un composant matériel ou un ensemble de composants matériels et logiciels, un composant logiciel correspondant lui-même à un ou plusieurs programmes ou sous-programmes d'ordinateur ou de manière plus générale à tout élément d'un programme apte à mettre en œuvre une fonction ou un ensemble de fonctions.

[0047] Sur toutes les figures du présent document, les éléments et étapes identiques sont désignés par une même référence numérique.

[0048] Selon un premier aspect, la présente technique se rapporte à un procédé de génération dynamique d'une transcription textuelle d'un flux audio diffusé en continu. Le principe général d'un tel procédé est illustré en relation avec la [Fig.1], dans un mode de réalisation particulier de la technique proposée. Ce procédé est mis en œuvre au sein d'un dispositif comprenant des moyens de réception d'un flux de données source comprenant ledit flux audio FA. Un tel flux de données source peut être constitué uniquement du flux audio FA en question, ou prendre par exemple la forme d'un flux vidéo comprenant au moins une piste audio formant le flux audio FA. Ce dispositif est typiquement un dispositif intermédiaire d'une chaine de diffusion en direct du flux de données source, permettant notamment par exemple, comme décrit par la suite, d'enrichir le flux de données source reçu en provenance d'un serveur de diffusion en direct avant de le relayer vers une pluralité de lecteurs destinés à le restituer à des utilisateurs finaux consommateurs du contenu associé.

[0049] Dans une étape 11, le flux audio FA reçu directement ou extrait d'un flux vidéo par le dispositif de génération dynamique est traité par un système de reconnaissance vocale, tel que par exemple du type « *Speech to Text* », NLP (« *Natural Language Processing* » en anglais), etc. Un tel système est intégré au dispositif de génération, ou bien de manière alternative connecté à ce dispositif via un réseau de communication. Plus particulièrement, le flux audio FA est transmis au système de reconnaissance vocale au fur et à mesure de sa réception, ce système délivrant en retour, par segment, un flux de données textuelles représentatif d'une première transcription textuelle du flux audio FA, dite transcription textuelle brute TTB en ce sens qu'elle n'a à ce stade fait l'objet d'aucune vérification.

[0050] Dans une étape 12, la transcription textuelle brute TTB est transmise par segment, i.e. au fur et à mesure de sa génération à l'étape 11, à un module de validation de transcription textuelle du dispositif de génération. Un tel module de validation met à disposition d'un utilisateur USR, par exemple via une application logicielle frontale (de l'anglais « *front-end* »), au moins une interface homme-machine IHM permettant de consulter en temps réel et au fil de l'eau, sur un écran d'affichage, la transcription textuelle brute TTB délivrée par le système de reconnaissance vocale. Via cette interface homme-machine IHM, l'utilisateur USR a ainsi la possibilité de valider la transcription textuelle, éventuellement après l'avoir modifiée sur le fond (par exemple à des fins de correction, de reformulation, de modération, etc.) et/ou sur la forme (par exemple pour y ajouter des retours à la ligne, des tirets, modifier la police ou la couleur du texte de manière à mieux différencier les locuteurs, etc.) grâce à des outils disponibles à cet effet au sein du module de validation.

[0051] Dans une étape 13, le dispositif de génération obtient ainsi par segment, au fur et à mesure de leur validation dans le module de validation, une nouvelle transcription textuelle du flux audio FA, dite transcription textuelle validée TTV en ce sens qu'elle a été revue et validée par un utilisateur.

[0052] On dispose ainsi d'une solution hybride, permettant d'obtenir une transcription textuelle de meilleure qualité que celle qui aurait été obtenue via un système automatique de reconnaissance de la parole utilisé seul, tout en étant moins coûteuse qu'une solution de vélotypie pure, dans la mesure où l'utilisateur en charge de valider la transcription n'a pas besoin d'avoir toutes les compétences d'un spécialiste vélotypiste car il a déjà accès, grâce à la technique proposée, à une base de transcription textuelle brute qu'il lui suffit d'adapter pour produire une retranscription fiable du flux audio diffusé en direct, le tout quasiment en temps-réel.

[0053] Selon une caractéristique particulière, la technique proposée offre également la possibilité, par exemple au moyen d'une option de paramétrage dédiée, de désactiver la validation par un humain de la transcription textuelle générée en étape 11 en sortie du

système de reconnaissance vocale. Une telle option de paramétrage, accessible par exemple au sein d'une interface d'administration du dispositif de génération dynamique selon la présente technique, permet notamment d'assurer la fourniture d'une transcription textuelle du flux audio diffusé en continu, y compris dans le cas où une personne normalement en charge de la validation de la transcription textuelle générée automatiquement s'avèrerait indisponible ou devrait s'absenter temporairement. Lorsqu'il est activé, un tel débrayage de la validation par un humain de la transcription textuelle peut prendre différentes formes. Dans une première implémentation, il peut par exemple consister en une désactivation pure et simple des étapes 12 et 13 précédemment décrites, qui ne sont alors plus mises en œuvre, au moins temporairement. De manière alternative, dans une deuxième implémentation, les étapes 12 et 13 restent mises en œuvre, mais la transcription textuelle générée par le système de reconnaissance vocale puis fournie au module de validation de transcription textuelle est automatiquement validée par ce module au bout d'un temps prédéterminé (par exemple de l'ordre d'une dizaine de secondes, éventuellement configurable au sein de l'interface d'administration mentionnée précédemment), lorsque aucune interaction d'un utilisateur avec l'interface homme-machine mise à disposition par ce module de validation n'est détectée durant ce laps de temps.

[0054] De manière optionnelle, dans une étape 14, la transcription textuelle validée TTV fait éventuellement l'objet d'une traduction, par un système de traduction textuelle automatique, dans au moins une autre langue que la langue d'origine associée au flux audio FA. Ce système de traduction automatique est intégré au dispositif de génération, ou bien de manière alternative connecté à ce dispositif via un réseau de communication. Une telle traduction est effectuée au fil de l'eau, le système de traduction automatique délivrant par segment au moins une transcription textuelle traduite TTT de la transcription textuelle validée TTV.

[0055] La transmission textuelle validée TTV et/ou la ou les transcriptions textuelles traduites TTT générées à la volée peuvent ensuite être exploitées pour différents usages, et notamment pour enrichir le flux de données audio ou vidéo source comme décrit par la suite dans un mode de réalisation particulier de la présente technique.

[0056] On décrit maintenant, en relation avec la [Fig.2], un exemple d'architecture fonctionnelle d'un dispositif 20 de génération dynamique d'une transcription textuelle d'un flux audio diffusé en continu, dans un mode de réalisation particulier de la technique proposée. Plus particulièrement, le dispositif illustré en [Fig.2] est notamment adapté pour la génération de sous-titres associés à un flux de données source diffusé en direct, prenant la forme d'un flux vidéo comprenant au moins une piste audio. Comme décrit par la suite, différents aspects complémentaires du procédé de génération dynamique d'une transcription textuelle sont mis en évidence au travers de cet exemple, dans

différents modes de réalisation particuliers de la technique proposée.

[0057] Dans cet exemple d'architecture, donné à titre illustratif et non limitatif, le dispositif 20 est divisé en deux blocs fonctionnels principaux – un bloc 21 de gestion du flux en direct et un bloc 22 de génération en direct d'au moins une transcription textuelle – comprenant chacun différents modules décrits par la suite.

[0058] Le bloc de gestion du flux en direct 21 reçoit en entrée, au niveau d'un module de manipulation de flux 211, un flux vidéo FVO émis en direct (flux vidéo original), par exemple en provenance d'un serveur de vidéos SRV_V. Le module de manipulation de flux 211 réalise alors d'une part une mise en mémoire tampon du flux vidéo original FVO, et extrait d'autre part, sous forme d'un flux audio FA, la piste audio associée au flux vidéo FVO. Comme détaillé par la suite, le flux vidéo original FVO est plus particulièrement placé dans la mémoire tampon pendant une temporisation d'une durée prédéterminée, typiquement de l'ordre de quelques secondes à quelques dizaines de secondes, mais généralement inférieure à une minute. Une telle temporisation permet de retarder légèrement le flux vidéo le temps que les opérations de transcriptions textuelles du flux vidéo original FVO soient effectuées. La durée de temporisation étant par ailleurs très faible, les conditions de diffusion restent assimilables à celles d'une diffusion en direct.

[0059] Le flux audio FA extrait est ensuite transmis au bloc 22 de génération en direct d'au moins une transcription textuelle, où il est reçu au niveau d'un module de gestion de l'audio 221.

[0060] Le module de gestion de l'audio 221 relaie ce flux audio FA à un moteur de reconnaissance vocale 222. Le module de gestion de l'audio 221 reçoit en retour, par segment, un flux de données textuelles correspondant au texte reconnu, au fur et à mesure du traitement du flux audio par le moteur de reconnaissance vocale 222. Selon une caractéristique particulière, chaque segment textuel ainsi reconnu et transmis au module de gestion de l'audio 221 par le moteur de reconnaissance vocale 222 comprend des informations d'horodatage de début et de fin du segment audio correspondant dans le flux vidéo original FVO.

[0061] Le module de gestion de l'audio 221 transmet ce flux de données textuelles (et les éventuelles informations d'horodatage associées) à un module de validation 223, où il est affiché à la volée dans une interface homme-machine de validation afin d'être rapidement contrôlé, éventuellement modifié et/ou modéré et/ou reformulé, puis validé, par segment, par un utilisateur USR, via des moyens idoines mis à disposition au sein du module de validation.

[0062] Les segments validés du flux de données textuelles sont renvoyés au module de gestion de l'audio 221. De manière optionnelle, comme déjà présenté en relation avec la [Fig.1], le module de gestion de l'audio 221 transmet le flux de données textuelles

validé à un moteur de traduction 224. Le module de gestion de l'audio 221 reçoit en retour, par segment, au moins un flux de données textuelles traduites, correspondant au texte validé traduit dans une langue différente de la langue d'origine du flux vidéo original. Chaque segment traduit reste associé aux informations d'horodatage de début et de fin du segment audio correspondant, dans sa langue d'origine, dans le flux vidéo original FVO.

[0063] À ce point, on dispose, au niveau du bloc de génération en direct d'au moins une transcription textuelle 22, du flux de données textuelles validées dans la langue d'origine du flux vidéo original, et éventuellement d'un ou plusieurs flux de données textuelles traduites dans d'autres langues.

[0064] Ces flux de données textuelles sont transmis au bloc de gestion du flux en direct 21, où ils sont reçus au fil de l'eau, par segment, au niveau d'un module d'enrichissement de flux 212. Au fur et à mesure de leur réception, dans un mode de réalisation particulier, le module d'enrichissement de flux 212 intègre alors à la volée ces flux de données textuelles au flux vidéo récupéré de la mémoire tampon où il a été placé par le module de manipulation de flux 211. Comme décrit précédemment, ce flux vidéo récupéré en mémoire tampon correspond plus particulièrement au flux vidéo original retardé (ou temporisé) d'une temporisation d'une durée prédéterminée au moins égale à la durée nécessaire pour générer le flux de données textuelles validées et éventuellement traduites. Selon une caractéristique particulière de la technique proposée, ces flux de données textuelles sont intégrés dans le flux vidéo temporisé, sous forme de métadonnées correspondant à leur langue respective, en tenant compte des informations d'horodatage disponibles pour chaque segment desdits flux de données. Par exemple, comme illustré en relation avec la [Fig.3], à chaque fois qu'une trame ou qu'un groupe de trames vidéo (trames vidéo 1, 2, 3, 4 ou 5 sur la [Fig.3]) est lu dans la mémoire tampon, un segment associé (typiquement un mot, par exemple un des quatre mots W1, W2, W3 et W4 constitutifs de la phrase « *Bonjour, comment allez-vous ?* », ou un groupe de mots) du flux de données textuelles obtenu en sortie de l'opération GEN_TT de génération dynamique de la transcription textuelle du flux vidéo original FVO selon la présente technique est intégré au flux vidéo temporisé, sous forme de métadonnées de sous-titres. Plus particulièrement, dans une opération de synchronisation SYNC, c'est le segment dont les informations d'horodatage correspondent à celles de la trame courante lue dans le flux vidéo temporisé qui est inséré et associé à cette trame dans le flux vidéo temporisé (la temporisation d'une durée DLY obtenue grâce à la mise en mémoire tampon du flux vidéo original FVO permettant d'assurer que ce segment du flux de données textuelles validées et éventuellement traduites est d'ores et déjà disponible au moment de la lecture de la trame vidéo correspondante du flux vidéo temporisé). De cette manière, on obtient un flux vidéo enrichi FVE

présentant une synchronisation parfaite entre transcription textuelle et trames vidéo.

[0065] Le flux vidéo enrichi FVE, qui n'est retardé que de quelques secondes par rapport au flux vidéo original, peut ensuite être diffusé dans une configuration de quasi-direct à destination de lecteurs vidéo PLY_V qui, s'ils sont compatibles avec les normes relatives au sous-titrage (ce qui est le cas de la majeure partie des lecteurs actuels), permettent au consommateur final d'afficher ou non les sous-titres, et, le cas échéant, de choisir la langue de sous-titrage (sous réserve que le flux de données textuelles traduites dans la langue choisie soit disponible). Grâce à la technique proposée, ces sous-titres apparaissent en outre sans aucun décalage temporel lors de la lecture du flux vidéo enrichi FVE.

[0066] En d'autres termes, de manière synthétique, si on se réfère aux pastilles étiquetées 1, 2, 3, 4 et A, B, C, D sur la [Fig.2] :

[0067] - le bloc de gestion du flux en direct 21 est en charge de : 1. Extraire un flux audio correspondant à la piste audio du flux vidéo en direct ; 2. Mettre en mémoire tampon le flux vidéo en direct original pour retarder la diffusion en direct ; 3. Rassembler les sous-titres générés ; 4. Enrichir le flux vidéo en direct « à la volée » en y intégrant les sous-titres, tout en assurant la synchronisation avec la parole ; 5. Fournir le flux vidéo enrichi à la diffusion avec le faible retard préalablement configuré ;

[0068] - le bloc de génération en direct d'au moins une transcription textuelle 22 est en charge de : A. Générer la reconnaissance vocale à partir du flux audio en direct, en utilisant par exemple un système de reconnaissance automatique de la parole (ASR) ; B. Fournir une interface utilisateur pour permettre facilement la modification, la modération et la validation en direct de la reconnaissance vocale, afin de générer des sous-titres validés ; C. Effectuer éventuellement une traduction des sous-titres dans une autre langue ; et D. Fournir un flux de sous-titres, éventuellement en plusieurs langues, à intégrer dans le flux vidéo en direct légèrement temporisé.

[0069] Ainsi, une telle mise en œuvre selon au moins un mode de réalisation particulier de la technique proposée permet :

[0070] - de ne pas partir d'une « page vierge » pour la génération de la transcription textuelle d'un flux audio diffusé en direct, en bénéficiant des résultats d'un premier traitement brut et automatisé de reconnaissance vocale ;

[0071] - de donner le temps à un être humain de passer rapidement en revue le texte brut reconnu et de lui offrir la possibilité de la modifier et de le valider facilement en temps réel, via une interface homme-machine dédiée ;

[0072] - d'intégrer automatiquement la transcription textuelle validée obtenue, par exemple sous la forme de sous-titres, dans le flux vidéo diffusé en quasi-direct, avec une qualité de sous-titrage et de synchronisation équivalente à celle qu'il serait possible d'obtenir pour un contenu vidéo préenregistré et destiné à être diffusé en différé (i.e. pas en

direct).

[0073] Comme déjà décrit précédemment, selon un autre aspect, la technique proposée se rapporte également à un dispositif de génération dynamique d'une transcription textuelle d'un flux audio diffusé en continu, apte à réaliser le procédé présenté dans l'un quelconque de ses modes de réalisation. Plus particulièrement, un tel dispositif selon la présente technique comprend :

[0074] - des moyens de traitement du flux audio, comprenant un système de reconnaissance vocale, délivrant une première transcription textuelle du flux audio, dite transcription textuelle brute ;

[0075] - des moyens de fourniture de la transcription textuelle brute à un module de validation comprenant au moins une interface homme-machine de validation de la transcription textuelle brute ;

[0076] - des moyens d'obtention en provenance du module de validation, d'une deuxième transcription textuelle du flux audio, dite transcription textuelle validée, établie via ladite interface homme-machine à partir de la transcription textuelle brute.

[0077] La [Fig.4] représente, de manière schématique et simplifiée, la structure d'un tel dispositif, dans un mode de réalisation particulier. Le dispositif intermédiaire selon la technique proposée comprend par exemple une mémoire 41 constituée d'une mémoire tampon M, une unité de traitement 42, équipée par exemple d'un microprocesseur μP, et pilotée par le programme d'ordinateur Pg 43, mettant en œuvre des étapes du procédé de génération dynamique d'une transcription textuelle d'un flux audio diffusé en continu, selon au moins un mode de réalisation de l'invention. À cette fin, le dispositif de génération comprend également au moins une interface de communication (par exemple une interface de communication Ethernet), lui permettant de recevoir et d'émettre des flux de données (vidéo, audio, textuelles) en provenance et à destination d'autres équipements présents dans le réseau de communication.

[0078] À l'initialisation, les instructions de code du programme d'ordinateur 43 sont chargées dans la mémoire tampon avant d'être exécutées par le processeur de l'unité de traitement 42. L'unité de traitement 42 reçoit en continu en entrée E un flux de données source comprenant un flux audio, en provenance d'un équipement d'une chaine de diffusion en direct, par exemple en provenance d'un serveur de diffusion en direct.

[0079] Le microprocesseur de l'unité de traitement 42 réalise alors les étapes du procédé de génération dynamique d'une transcription textuelle du flux audio diffusé en continu, selon les instructions du programme d'ordinateur 43. Plus particulièrement, le flux audio est traité dans un premier temps par un système de reconnaissance vocale, et le flux de données textuelles brutes résultant est traité dans un deuxième temps par un opérateur humain au moyen d'une interface homme-machine dédiée délivrée par le

programme d'ordinateur 43, permettant ainsi d'obtenir en temps-réel ou quasi-temps-réel en sortie S une transcription textuelle validée du flux audio.

Revendications

[Revendication 1] Procédé de génération dynamique d'une transcription textuelle d'un flux audio (FA) diffusé en continu, mis en œuvre au niveau d'un dispositif recevant ledit flux audio, ledit procédé étant caractérisé en ce qu'il comprend les étapes suivantes :
- traitement (11) dudit flux audio (FA) par un système de reconnaissance vocale, délivrant une première transcription textuelle dudit flux audio (TTB) ;
- fourniture (12) de ladite première transcription textuelle (TTB) à un module de validation de transcription textuelle comprenant au moins une interface homme-machine (IHM) de validation de ladite première transcription textuelle ;
- obtention (13), en provenance dudit module de validation, d'une deuxième transcription textuelle dudit flux audio, dite transcription textuelle validée (TTV).

[Revendication 2] Procédé selon la revendication 1, caractérisé en ce qu'il comprend en outre une étape de traduction (14) de ladite transcription textuelle validée (TTV), délivrant au moins une troisième transcription textuelle dudit flux audio dans une langue autre qu'une langue d'origine associée audit flux audio, dite transcription textuelle traduite (TTT).

[Revendication 3] Procédé selon la revendication 2, caractérisé en ce que ledit flux audio correspond à une piste audio d'un flux vidéo source diffusé en continu.

[Revendication 4] Procédé selon la revendication 3, caractérisé en ce qu'il comprend en outre les étapes suivantes :
- mise en mémoire tampon dudit flux vidéo source pendant une temporisation d'une durée prédéterminée, délivrant un flux vidéo temporisé ;
- intégration à la volée de ladite transcription textuelle validée et/ou de ladite au moins une transcription textuelle traduite au sein dudit flux vidéo temporisé, délivrant un flux vidéo temporisé enrichi.

[Revendication 5] Procédé selon la revendication 4, caractérisé en ce que ladite durée prédéterminée est inférieure à une minute.

[Revendication 6] Procédé selon la revendication 4, caractérisé en ce que ladite intégration comprend l'ajout à la volée, audit flux vidéo temporisé, de métadonnées comprenant ladite transcription textuelle validée et/ou ladite au moins une transcription textuelle traduite.

[Revendication 7] Procédé selon la revendication 6, caractérisé en ce que ledit ajout de métadonnées audit flux vidéo temporisé tient compte de données

d'horodatage associées au traitement mis en œuvre par ledit système de reconnaissance vocale.

[Revendication 8] Dispositif de génération dynamique d'une transcription textuelle d'un flux audio diffusé en continu, ledit dispositif étant caractérisé en ce qu'il comprend :

- des moyens de traitement dudit flux audio, comprenant un système de reconnaissance vocale, délivrant une première transcription textuelle dudit flux audio ;
- des moyens de fourniture de ladite première transcription textuelle à un module de validation de transcription textuelle comprenant au moins une interface homme-machine de validation de ladite première transcription textuelle ;
- des moyens d'obtention, en provenance dudit module de validation, d'une deuxième transcription textuelle dudit flux audio, dite transcription textuelle validée.

[Revendication 9] Produit programme d'ordinateur téléchargeable depuis un réseau de communication et/ou stocké sur un support lisible par ordinateur et/ou exécutable par un microprocesseur, caractérisé en ce qu'il comprend des instructions de code de programme pour l'exécution d'un procédé selon l'une quelconque des revendications 1 à 7, lorsqu'il est exécuté par un ordinateur.

[Revendication 10] Support d'enregistrement lisible par ordinateur comprenant des instructions de code de programme qui, lorsqu'elles sont exécutées par un ordinateur, conduisent celui-ci à mettre en œuvre le procédé selon l'une quelconque des revendications 1 à 7.

[Fig. 1]

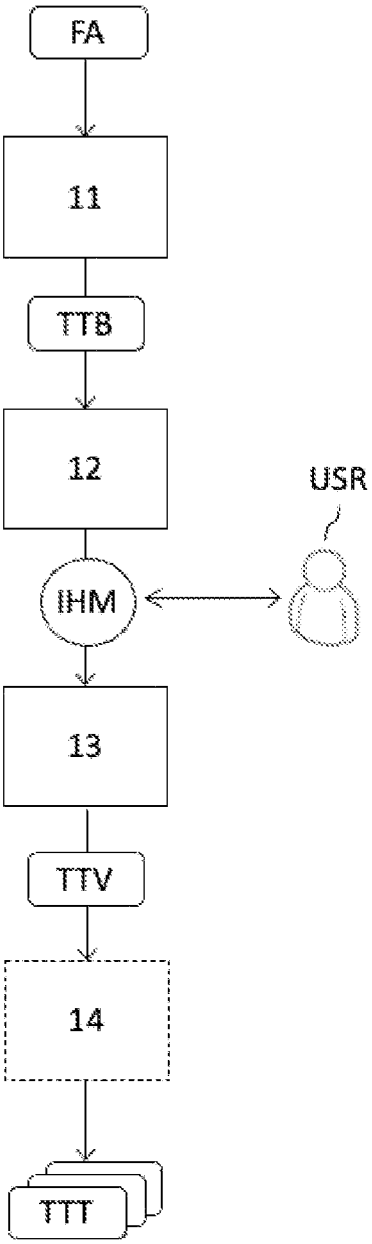


Fig. 1

[Fig. 2]

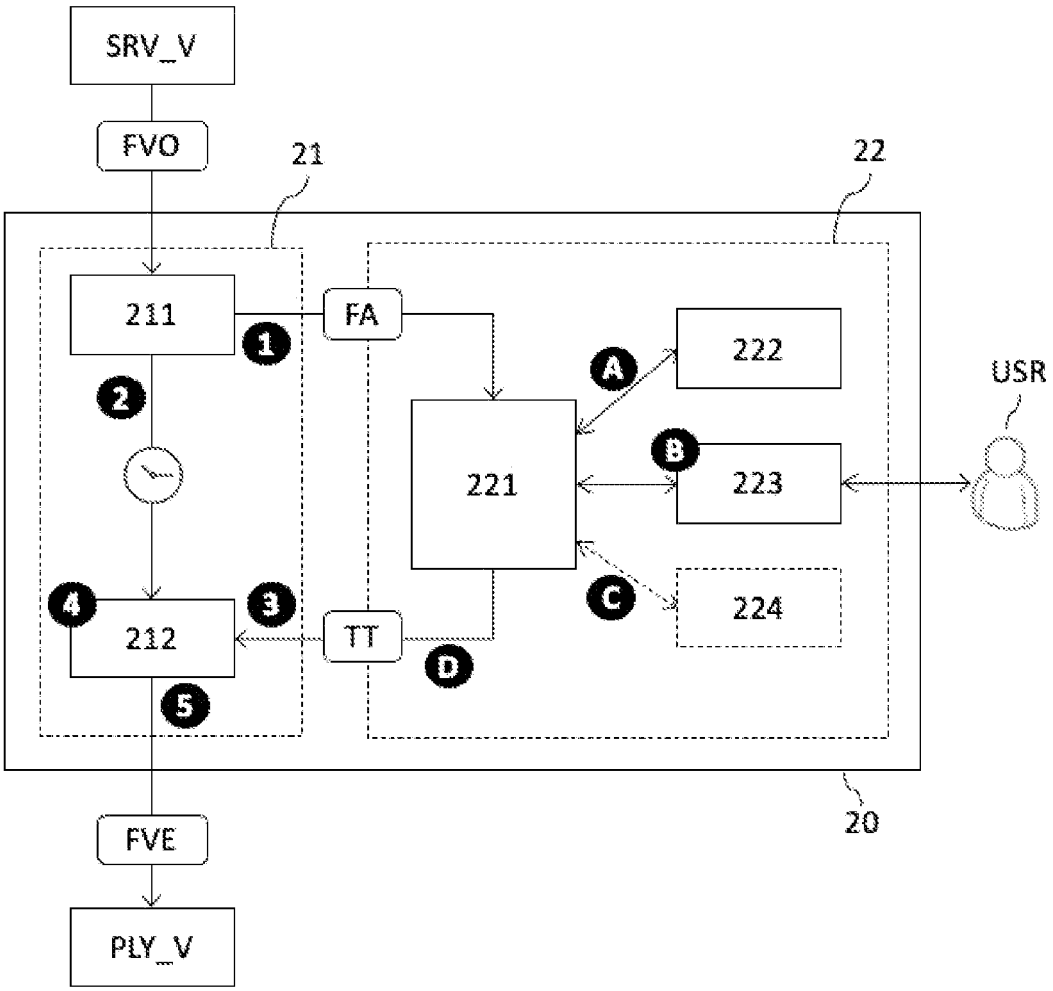


Fig. 2

[Fig. 3]

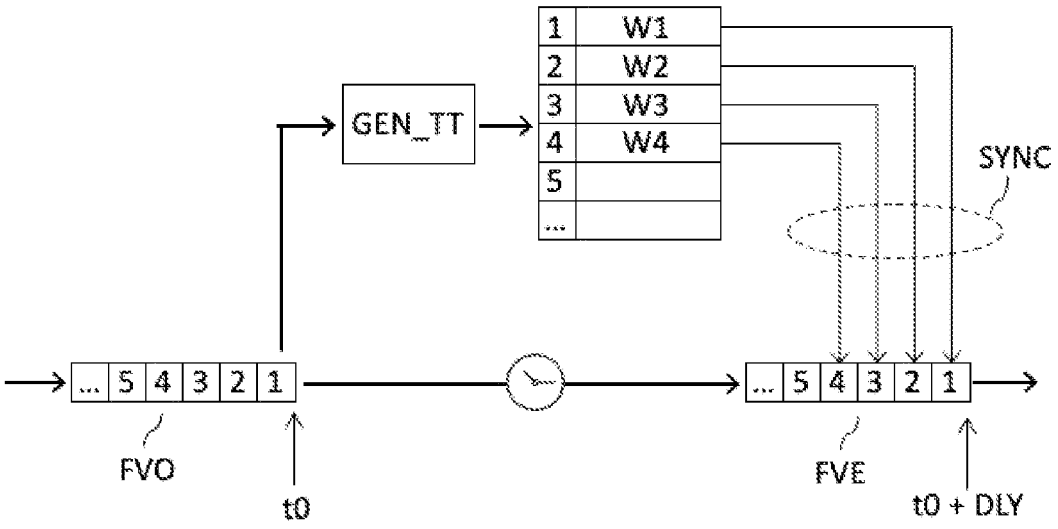


Fig. 3

[Fig. 4]

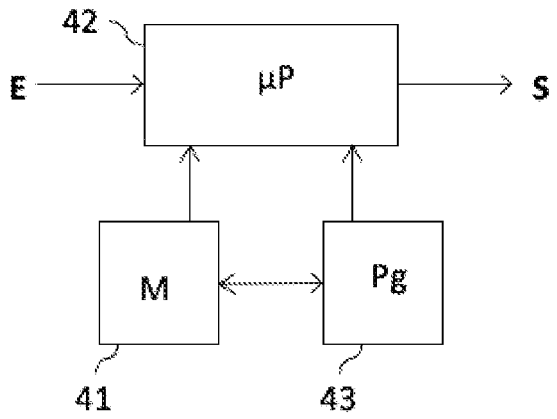


Fig. 4

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 909585
FR 2206706

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 2017/085696 A1 (ABKAIROV NIKOLAY [US]) 23 mars 2017 (2017-03-23) * alinéa [0001] - alinéa [0030] * * alinéa [0055] - alinéa [0056] * -----	1-10	H04N21/43 H04N21/439
X	US 2015/098018 A1 (STARLING MICHAEL IRVING [US] ET AL) 9 avril 2015 (2015-04-09) * alinéa [0005] - alinéa [0025] * * alinéa [0041] - alinéa [0047] * -----	1-10	
X	US 2021/074277 A1 (LEWIS WILLIAM DUNCAN [US]) 11 mars 2021 (2021-03-11) * alinéa [0002] - alinéa [0003] * * alinéa [0016] - alinéa [0038] * -----	1-10	
A	MERCEDES DE CASTRO ET AL: "Real-time subtitle synchronization in live television programs", BROADBAND MULTIMEDIA SYSTEMS AND BROADCASTING (BMSB), 2011 IEEE INTERNATIONAL SYMPOSIUM ON, IEEE, 8 juin 2011 (2011-06-08), pages 1-6, XP031894608, DOI: 10.1109/BMSB.2011.5954889 ISBN: 978-1-61284-121-2 * page 1, colonne 2, ligne 1 - ligne 9 * * page 5, colonne 1, ligne dernière - colonne 2, ligne troisième * -----	1-10	DOMAINES TECHNIQUES RECHERCHÉS (IPC) H04N G06F
X	US 2018/144747 A1 (SKARBOVSKY EVGENY [US] ET AL) 24 mai 2018 (2018-05-24) * alinéa [0041] - alinéa [0045] * -----	1-3,6-10	

Date d'achèvement de la recherche	Examinateur
31 janvier 2023	Vaquero, Raquel

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 2206706 FA 909585

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 31-01-2023
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 2017085696 A1	23-03-2017	CN 108028042 A	11-05-2018
		EP 3347895 A1	18-07-2018
		US 2017085696 A1	23-03-2017
		WO 2017048588 A1	23-03-2017
US 2015098018 A1	09-04-2015	AUCUN	
US 2021074277 A1	11-03-2021	EP 4026119 A1	13-07-2022
		US 2021074277 A1	11-03-2021
		WO 2021045828 A1	11-03-2021
US 2018144747 A1	24-05-2018	AUCUN	

EPO FORM P0465