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

(11) **EP 3 828 886 A1**

(12) **DEMANDE DE BREVET EUROPEEN**

(43) Date de publication:
02.06.2021 Bulletin 2021/22

(51) Int Cl.:
G10L 21/0272 *(2013.01)*

(21) Numéro de dépôt: **20209511.3**

(22) Date de dépôt: **24.11.2020**

(84) Etats contractants désignés:
AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO RS SE SI SK SM TR
Etats d'extension désignés:
BA ME
KH MA MD TN

(30) Priorité: **27.11.2019 FR 1913283**

(71) Demandeur: **THALES**
92400 Courbevoie (FR)

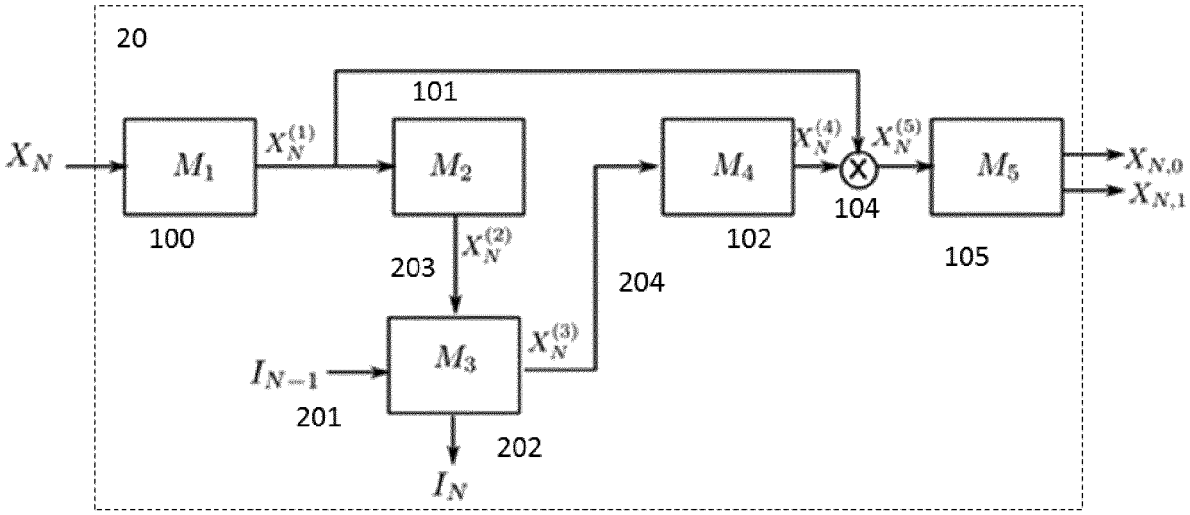
(72) Inventeurs:
- **MATHIEU, Félix**
  91767 PALAISEAU CEDEX (FR)
- **COURTAT, Thomas**
  91767 PALAISEAU CEDEX (FR)
- **CAPMAN, François**
  92622 GENNEVILLIERS CEDEX (FR)
- **SAUSSET, François**
  91767 PALAISEAU CEDEX (FR)
- **ACHECHE, Shaheen**
  91767 PALAISEAU CEDEX (FR)

(74) Mandataire: **Marks & Clerk France**
Immeuble "Visium"
22, avenue Aristide Briand
94117 Arcueil Cedex (FR)

(54) **PROCEDE ET SYSTEME POUR SEPARER DANS UN FLUX AUDIO LA COMPOSANTE VOIX ET LA COMPOSANTE BRUIT**

(57) L'invention concerne un procédé et un système pour séparer en temps réel dans un flux audio la composante voix et la composante bruit.

[Fig. 3]



EP 3 828 886 A1

Description

[0001] L'invention concerne un procédé et un système permettant de séparer, en temps réel dans un flux audio, la partie du flux associée à une voix ou à de la parole, d'une autre partie du flux contenant les bruits.

[0002] L'invention trouve son application dans un contexte où une ou plusieurs personnes parlent dans un environnement bruité (brouhaha, bruit de moteur, ventilation, etc.). Le signal de la parole superposé aux signaux bruyants est numérisé dans un flux audio par un capteur sonore.

[0003] L'invention concerne aussi un procédé et un système pour rehausser un signal de voix en temps réel dans un flux audio à partir d'un procédé de séparation de sources audio en temps différé.

[0004] L'état de l'art connu du demandeur se divise en deux catégories, les approches dites classiques et les approches possibles par l'intelligence artificielle connue sous la dénomination anglo-saxonne de « deep learning ».

[0005] Dans l'approche de « deep learning », des approches traitent directement du problème de séparation voix/bruit de fond, d'autres concernent la séparation signal/signal, voix/voix.

[0006] La demande de brevet US 20190066713 divulgue un procédé consistant à obtenir, par un dispositif, un signal sonore combiné pour des signaux combinés provenant de multiples sources sonores dans une zone dans laquelle se trouve une personne. Le traitement mis en œuvre fait appel à des réseaux de neurones profonds.

[0007] Un exemple de procédé pour séparer plusieurs voix dans un signal audio selon l'art antérieur comporte les étapes décrites ci-après et non représentées pour des raisons de simplification. Le signal audio entrant est noté X, il a pour longueur L. Le signal est transmis à un encodeur M_1 qui transforme X en un tenseur $X^{(1)}$de dimensions $F \times T$ où T est un diviseur de $\underline{L}$ et $\underline{F}$ un nombre de filtres donné par le concepteur. L'encodeur M_1 consiste en une Convolution 1D à F filtres. Les coefficients des noyaux de convolution sont réglés lors d'une phase d'apprentissage. Le tenseur est transmis d'une part à un multiplicateur pour une utilisation future et d'autre part à un module de séparation. Le module de séparation est divisé en deux sous-modules M_2 et M_4. Le premier sous-module M_2 transforme le tenseur $X^{(1)}$ en un tenseur $X^{(2)}$ de dimensions $F \times T$. Le premier sous-module M_2 est constitué d'une couche de normalisation, une convolution 1x1 et un empilement de modules 1D-Conv connus de l'art antérieur et dont les paramètres sont réglés lors d'une phase d'apprentissage.

[0008] Le deuxième sous-module M_4 transforme $X^{(2)}$ en $X^{(4)}$ tenseur de dimensions $2F \times T$. Pour cela, le deuxième sous-module M_4 enchaîne une non-linéarité, une convolution 1x1 et une fonction sigmoïde. Les coefficients de la convolution 1x1 sont réglés lors d'une phase d'apprentissage.

[0009] $X^{(1)}$ est concaténé à lui-même pour former un tenseur de dimensions 2F x T qui est multiplié à $X^{(4)}$ pour former $X^{(5)}$.

[0010] Le module M_5 prend pour entrée $X^{(5)}$ et donne en sortie deux signaux de longueur $\underline{L}$ au moyen d'une déconvolution 1D dont les paramètres sont réglés lors d'une phase d'apprentissage.

[0011] Les paramètres numériques définissant les traitements des différents modules sont obtenus dans une phase préalable d'apprentissage sur une base de données.

[0012] En remplaçant une des voix par du bruit, il est immédiat d'utiliser les méthodes décrites dans l'état de l'art pour séparer la voix du bruit de fond dans un signal audio et, en conservant uniquement la sortie contenant le signal de voix, de rehausser la voix d'un signal bruité.

[0013] La figure 1 illustre une application à la séparation de signaux de différents types, en séparant le canal voix et le canal bruit.

[0014] Tel que décrit, l'état de l'art ne permet pas directement le traitement en temps réel d'un flux audio.

[0015] Le document de Mimilakis Stylianos Ioannis et al, intitulé « A recurrent encoder-decoder approach with skip-filtering connections for monaural singing voice separation », du 25 septembre 2017, pages 1-6, XP 033263882, divulgue un procédé permettant de séparer la voix d'un fond musical.

[0016] Dans le domaine technique du « deep learning », les données sont représentées sous forme de tenseurs. Les données sont modifiées par une succession de modules. En sortie de chaque module, les données sont projetées dans un espace abstrait défini en général par ses dimensions.

[0017] Pour ce faire la présente invention met en œuvre les traitements suivants :

[0018] Le signal (flux d'entrée X) est découpé en N trames de longueur L, avec X_N la nième trame. Le procédé exécute les traitements suivants :

[0019] La trame X_N est encodée par un réseau de convolutions 1D. Le résultat est un tenseur $X_N^{(1)}$ de dimensions F x T avec F le nombre de filtres donné par le concepteur,

[0020] T un diviseur de L dépendant de la taille des filtres F, 100. Le résultat $X_N^{(1)}$ est ensuite transformé par un module M_2, 101. Le résultat $X_N^{(2)}$ est un tenseur de dimensions F x T. Le module M_4 estime, 103, à partir de $X_N^{(2)}$, un tenseur $X_N^{(4)}$ de dimensions 2F x T. $X_N^{(1)}$ est concaténé à lui-même, 104, pour former un tenseur de dimensions 2F x T qui est multiplié à $X_N^{(4)}$

pour former $X_N^{(5)}$. Le module M_5 à partir de $X_N^{(5)}$ produit un tenseur de dimension 2 x T, 105, à partir duquel on obtient deux sorties de dimensions 1 x T $X_{N,0}$ et $X_{N,1}$ qui sont respectivement le canal voix et le canal bruit.

[0021] Ces étapes sont réitérées sur chaque nouvelle trame. Les paramètres sont appris sur une base de données de sons. L'inconvénient de ce procédé est qu'il n'utilise pas les informations des trames précédentes pour traiter la trame courante. Ceci entraîne notamment une qualité dégradée et une forte latence dans les traitements, du fait de la durée des trames.

[0022] L'un des objectifs de la présente invention est d'offrir un procédé et un dispositif permettant de séparer, en temps réel, des voix du bruit de fond dans un flux audio, ou débruitage de la voix dans un flux audio, notamment en tenant compte des informations issues des trames précédentes. Ceci permet d'améliorer les performances et la latence de traitement. Le procédé permet ainsi la propagation de « l'information globale » sur le signal, sa mise à jour et son exploitation de trame en trame.

[0023] L'invention concerne un procédé pour séparer en temps réel de la voix du bruit dans un signal audio reçu sur un récepteur équipé d'un capteur audio caractérisé en ce qu'il comporte au moins les étapes suivantes :

- On sépare le flux audio reçu en N trames X_N,
- Pour chaque trame X_N on associe un tenseur contenant des informations sur l'ensemble du flux audio,
- On transmet la trame X_N à un premier module M_1 qui génère un signal $X_N^{(1)}$,
- Le tenseur I_{N-1} obtenu lors de l'étape précédente pour le traitement de la trame X_N - 1 est transmis à un module M_3,
- Le module M_3 prend en entrée un signal $X_N^{(2)}$, résultat de la transformation du signal $X_N^{(1)}$ par un module M_2 et réalise la concaténation de $X_N^{(2)}$ et I_N afin de générer un signal $X_N^{(3)}$ de dimension 2F x T,
- Le signal $X_N^{(3)}$ est transmis à un module M_4 afin de générer un signal $X_N^{(4)}$ qui est combiné avec le signal $X_N^{(1)}$,
- Le signal résultant de la combinaison est décodé par un décodeur M_5 afin de générer un premier signal de voix $X_{N,0}$ et un deuxième signal $X_{N,1}$.

[0024] Pour traiter une trame N on suppose que la trame N - 1 a été traitée précédemment et que les quantités résultant de ce traitement ont été stockées. Pour la trame 0, I_0 est fixé arbitrairement par exemple il est identiquement nul.

[0025] L'invention concerne aussi un dispositif pour séparer de la voix du bruit dans un signal audio reçu sur un récepteur équipé d'un capteur audio caractérisé en ce qu'il comporte au moins les éléments suivants :

- Un premier module M_1 recevant des trames d'un signal contenant de la voix et du bruit,
- Le premier module à une sortie reliée à un deuxième module M_2 configuré pour générer un signal transmis à un troisième module M_3 qui reçoit une valeur de tenseur associée à une trame précédente X_N - 1 pour générer un tenseur I_N associé à la trame courante et un signal $X_N^{(3)}$ de dimension 2F x T,

[0026] Le module M_3 inséré entre le module M_2 et le module M_4 prend en entrée un tenseur homogène en dimensions à celui fourni en sortie du module M_2 et fournit en sortie un tenseur homogène en dimensions à celui que prend en entrée le module M_4. Une entrée I_N - 1 supplémentaire est fournie en entrée du module M_3 pour le traitement de la trame numéro N et le module M_3 fournit en sortie additionnelle le tenseur I_N.

- Un module M_4 qui combine le signal $X_N^{(3)}$ et le signal $X_N^{(1)}$ afin de générer un signal $X_N^{(4)}$,
- Un décodeur M_5 configuré pour générer un premier signal de voix $X_{N,0}$ et un deuxième signal de bruit $X_{N,1}$ à partir du signal $X_N^{(4)}$.

[0027] D'autres caractéristiques, détails et avantages de l'invention ressortiront à la lecture de la description faite en référence aux dessins annexés donnés à titre d'exemple non limitatifs et qui représentent, respectivement :

[Fig.1], une illustration de l'art antérieur,
[Fig.2], un exemple de système permettant la mise en œuvre du procédé selon l'invention,
[Fig.3] une illustration des étapes mises en œuvre par le procédé selon l'invention.

[0028] La figure 2 illustre un exemple de dispositif permettant la mise en œuvre du procédé selon l'invention.

[0029] Le signal dont il faut extraire (séparer) la ou les voix du bruit contenu dans le flux audio est reçu sur un capteur audio 10. Le capteur audio est relié à un ensemble d'équipements ou modules Hardware 20 configurés pour séparer la voix du bruit qui seront détaillés à la figure 3.

[0030] La figure 3 illustre une première variante de réalisation pour séparer une voix du bruit dans un signal audio, les traitements étant effectués au niveau de l'ensemble 20. Cette séparation est réalisée en temps réel. Les modules similaires au schéma de la figure 1 portent les mêmes références. L'ensemble comprend en plus un module M_3 dont la fonction est détaillée ci-après.

[0031] Le signal audio reçu sur le capteur est lors d'une première étape séparé en N trames X_1.....X_N. A chaque trame X_N est associé un tenseur I_N qui est de dimension constante, indépendante de l'indice de la trame. Le procédé va mettre à jour la valeur du tenseur I_N de trame en trame et l'utilisation jointe de X_N et I_N pour estimer $X_{N,0}$ et $X_{N,1}$.

[0032] La trame X_N est transmise à un premier module M_1, 100, qui génère un signal $X_N^{(1)}$. Le tenseur I_{N-1} obtenu lors de l'étape précédente pour le traitement de la trame X_N - 1 est transmis dans un module M_3, 201.

[0033] M_3 génère un tenseur I_N, 202, qui sera utilisé lors du traitement de la trame X_{N+1}.

[0034] Le codeur M_3 prend en entrée un signal $X_N^{(2)}$, 203, résultat de la transformation du signal $X_N^{(1)}$ par un module M_2 et réalise la concaténation de $X_N^{(2)}$ et I_N, afin de générer un signal $X_N^{(3)}$ de dimension 2F x T,

$$M_{3:}(X_N^{(2)}, I_{N-1}) -> (X_N^{(3)}, I_N).$$

[0035] Le signal $X_N^{(3)}$, 204, est transmis à un module M_4 afin de générer un signal $X_N^{(4)}$ qui est combiné, 104, avec le signal $X_N^{(1)}$, le signal résultant de la combinaison est décodé par un décodeur M_5, 105, afin de générer un premier signal de voix $X_{N,0}$ et un deuxième signal de bruit $X_{N,1}$.

[0036] Dans un mode de réalisation, les étapes mises en œuvre par le procédé selon l'invention sont les suivantes :

[0037] Pour tout N, I_N est de dimension F x F

[0038] A_N est un tenseur F x F défini par

$$A_N = \left(\frac{X_N^{(2)} \cdot \left(X_N^{(2)}\right)^t}{\sqrt{T}} \right)$$

a. $X_N^{(2)} \cdot \left(X_N^{(2)}\right)^t$ est le produit matriciel de $X_N^{(2)}$ et de sa transposée

$I_N = I_{N-1} + \lambda(A_N - I_{N-1})$ avec λ un facteur de gain 0 et 1 donné par l'utilisateur

$B_N = Softmax(I_{N-1})$

a. La fonction softmax est classique en machine learning ; à un vecteur de K nombres, (v_1 ...v_K) elle associe un vecteur de K nombre (w_1 ... w_K) avec

$$w_k = \frac{\exp(v_k)}{\sum_{l=1}^{K} \exp(v_l)},$$

pour tout

b. Pour calculer B_N, la fonction softmax est appliquée indépendamment à toutes les lignes de I_N,

$C_N = B_N \cdot X_N^{(2)}$ est le produit matriciel entre B_N et $X_N^{(2)}$; ses dimensions sont $F \times T$, $X_N^{(3)}$ est de dimension 2F x T, c'est la concaténation de $X_N^{(2)}$ et C_N.

[0039] Le procédé et le dispositif selon l'invention permettent une séparation en temps réel de la voix du bruit dans un signal audio reçu sur un capteur en temps réel et sans dégrader les paramètres propres à la voix.

[0040] Les paramètres numériques définissant les traitements des différents modules sont réglés dans une phase préalable d'apprentissage sur une base de données.

[0041] L'invention permet un fonctionnement en temps réel avec un compromis latence/qualité contrôlable, de ne pas dégrader le signal audio qui ne contient pas de bruit, et permet de rehausser le bruit dans un signal ne contenant pas de paroles (de voix).

[0042] Le procédé permet notamment de prétraiter le signal audio de la parole pour améliorer la qualité de briques de traitement / valorisation de la voix (compression, analyse).

[0043] L'ajout dans la chaîne de traitement d'un module M_3 permet d'améliorer la qualité de mise en place d'une stratégie trame par trame pour la mise en temps réel des traitements.

Revendications

1. Procédé pour séparer, en temps réel, de la voix du

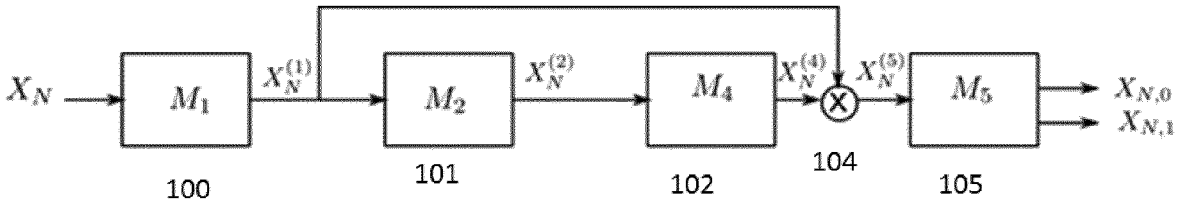
bruit dans un signal audio reçu sur un récepteur équipé d'un capteur audio **caractérisé en ce qu'**il comporte au moins les étapes suivantes :

- On sépare le flux audio reçu en N trames X_N,
- Pour chaque trame X_N on associe un tenseur contenant des informations sur l'ensemble du flux audio,
- On transmet la trame X_N à un premier module M_1, (100), qui génère un signal $X_N^{(1)}$,
- Le tenseur I_{N-1} obtenu lors de l'étape précédente pour le traitement de la trame X_N - 1 est transmis à un module M_3, (201),
- Le module M_3 prend en entrée un signal $X_N^{(2)}$, (203), résultat de la transformation du signal $X_N^{(1)}$ par un module M_2 et réalise la caténation de $X_N^{(2)}$ et I_N, afin de générer un signal $X_N^{(3)}$ de dimension 2F x T,
- Le signal $X_N^{(3)}$, (204), est transmis à un module M_4 afin de générer un signal $X_N^{(4)}$ qui est combiné, (104), avec le signal $X_N^{(1)}$,
- le signal résultant de la combinaison est décodé par un décodeur M_5, (105), afin de générer un premier signal de voix $X_{N,0}$ et un deuxième signal $X_{N,1}$.

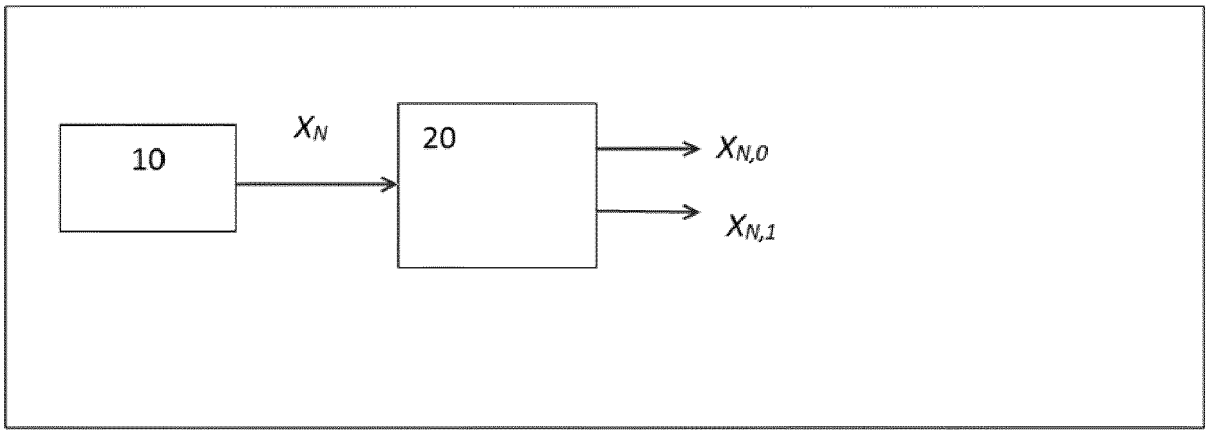
2. Dispositif pour séparer, en temps réel, de la voix du bruit dans un signal audio reçu sur un récepteur équipé d'un capteur audio **caractérisé en ce qu'**il comporte au moins les éléments suivants :

- Un premier module M_1 recevant des trames d'un signal contenant de la voix et du bruit,
- Le premier module à une sortie reliée à un deuxième module M_2 configuré pour générer un signal transmis à un troisième module M_3 qui reçoit une valeur de tenseur associée à une trame précédente X_N - 1 pour générer un tenseur I_N associé à la trame courante et un signal $X_N^{(3)}$ de dimension 2F x T,
- Un module M_4 qui combine le signal $X_N^{(3)}$ et le signal $X_N^{(1)}$ afin de générer un signal $X_N^{(4)}$,
- Un module (104) qui combine le signal $X_N^{(4)}$ avec le signal $X_N^{(1)}$ afin de générer un signal $X_N^{(5)}$
- Un décodeur M_5, (105) configuré pour générer un premier signal de voix $X_{N,0}$ et un deuxième signal $X_{N,1}$ à partir du signal $X_N^{(5)}$.

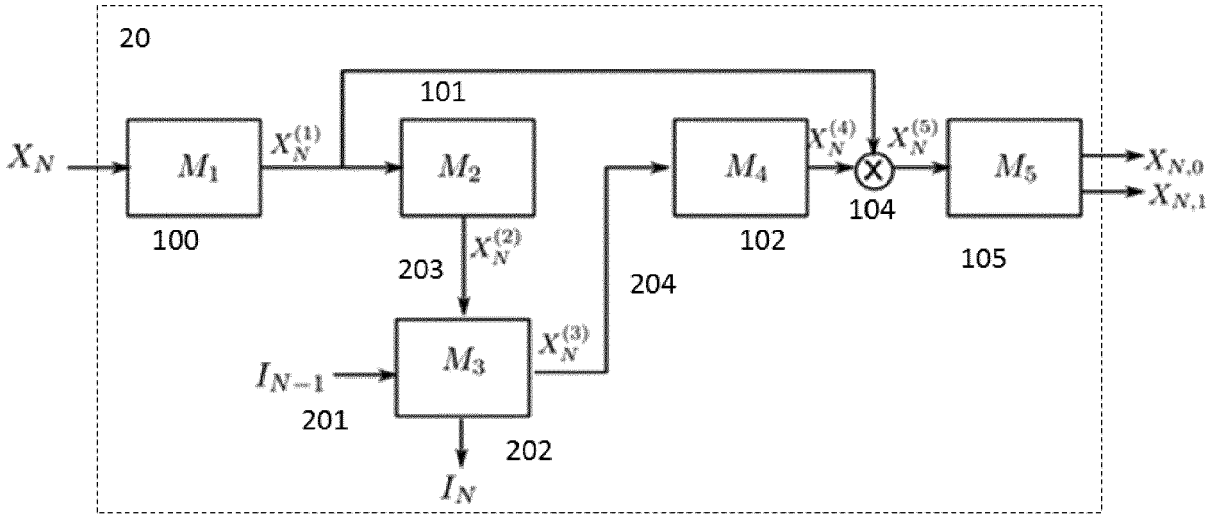
[Fig. 1]



[Fig. 2]



[Fig. 3]



Europäisches Patentamt
European Patent Office
Office européen des brevets

RAPPORT DE RECHERCHE EUROPEENNE

Numéro de la demande
EP 20 20 9511

DOCUMENTS CONSIDERES COMME PERTINENTS

Catégorie	Citation du document avec indication, en cas de besoin, des parties pertinentes	Revendication concernée	CLASSEMENT DE LA DEMANDE (IPC)
X	MIMILAKIS STYLIANOS IOANNIS ET AL: "A recurrent encoder-decoder approach with skip-filtering connections for monaural singing voice separation", 2017 IEEE 27TH INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING (MLSP), IEEE, 25 septembre 2017 (2017-09-25), pages 1-6, XP033263882, DOI: 10.1109/MLSP.2017.8168117 * abrégé * * figure 1 * * section 2. * * section 3.1., lignes 1-3. * -----	1,2	INV. G10L21/0272
A,D	US 2019/066713 A1 (MESGARANI NIMA [US] ET AL) 28 février 2019 (2019-02-28) * figure 14 * -----	1,2	
A	STEPHENSON CORY ET AL: "Monaural speaker separation using source-contrastive estimation", 2017 IEEE INTERNATIONAL WORKSHOP ON SIGNAL PROCESSING SYSTEMS (SIPS), IEEE, 3 octobre 2017 (2017-10-03), pages 1-6, XP033257056, DOI: 10.1109/SIPS.2017.8110005 * figure 2 * -----	1,2	DOMAINES TECHNIQUES RECHERCHES (IPC) G10L

Le présent rapport a été établi pour toutes les revendications

Lieu de la recherche	Date d'achèvement de la recherche	Examinateur
Munich	18 février 2021	Chétry, Nicolas

CATEGORIE DES DOCUMENTS CITES

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 antérieur, mais publié à la date de dépôt ou après cette date
D : cité dans la demande
L : cité pour d'autres raisons
.....
& : membre de la même famille, document correspondant

EPO FORM 1503 03.82 (P04C02)

ANNEXE AU RAPPORT DE RECHERCHE EUROPEENNE RELATIF A LA DEMANDE DE BREVET EUROPEEN NO. EP 20 20 9511

La présente annexe indique les membres de la famille de brevets relatifs aux documents brevets cités dans le rapport de recherche européenne visé ci-dessus.
Lesdits members sont contenus au fichier informatique de l'Office européen des brevets à la date du
Les renseignements fournis sont donnés à titre indicatif et n'engagent pas la responsabilité de l'Office européen des brevets.

18-02-2021

Document brevet cité au rapport de recherche	Date de publication	Membre(s) de la famille de brevet(s)	Date de publication
US 2019066713 A1	28-02-2019	AUCUN	

EPO FORM P0460

Pour tout renseignement concernant cette annexe : voir Journal Officiel de l'Office européen des brevets, No.12/82

RÉFÉRENCES CITÉES DANS LA DESCRIPTION

Cette liste de références citées par le demandeur vise uniquement à aider le lecteur et ne fait pas partie du document de brevet européen. Même si le plus grand soin a été accordé à sa conception, des erreurs ou des omissions ne peuvent être exclues et l'OEB décline toute responsabilité à cet égard.

Documents brevets cités dans la description

- US 20190066713 A **[0006]**

Littérature non-brevet citée dans la description

- **MIMILAKIS STYLIANOS LOANNIS et al.** *A recurrent encoder-decoder approach with skip-filtering connections for monaural singing voice separation,* 25 Septembre 2017, 1-6 **[0015]**