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

(11) **EP 3 040 989 A1**

(12) **DEMANDE DE BREVET EUROPEEN**

(43) Date de publication:
06.07.2016 Bulletin 2016/27

(51) Int Cl.:
G10L 21/0208 *(2013.01)*

(21) Numéro de dépôt: **15198713.8**

(22) Date de dépôt: **09.12.2015**

(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
Etats de validation désignés:
MA MD

(30) Priorité: **31.12.2014 FR 1463482**

(71) Demandeur: **Audionamix**
75010 Paris (FR)

(72) Inventeur: **HENNEQUIN, Romain**
75014 PARIS (FR)

(74) Mandataire: **Blot, Philippe Robert Emile**
Cabinet Lavoix
2, place d'Estienne d'Orves
75441 Paris Cedex 09 (FR)

(54) PROCÉDÉ DE SÉPARATION AMÉLIORÉ ET PRODUIT PROGRAMME D'ORDINATEUR

(57) Procédé consistant à séparer, dans un signal de mélange ($w(t)$), une contribution spécifique pure $x(t)$ et une contribution de fond sonore $z(t)$ en utilisant un spectrogramme de modélisation du signal de mélange $\hat{V}$ correspondant à la somme d'un spectrogramme d'une contribution spécifique réverbérée $\hat{V}^{rev,y}$ et d'un spectrogramme de la contribution de fond sonore $\hat{V}^{z}$, le spectrogramme de la contribution spécifique réverbérée dépendant du spectrogramme de la contribution pure $\hat{V}^{x}$ selon le modèle :

$$\hat{V}_{f,t}^{rev.y} = \sum_{\tau=1}^{T} \hat{V}_{f,t-\tau+1}^{x} R_{f,\tau}$$

où R est une matrice de réverbération, f est un pas de fréquence, t est un pas de temps, et τ un entier entre 1 et T ; et en minimisant une fonction de coût (C) entre le spectrogramme du signal de mélange et le spectrogramme de modélisation du signal de mélange.

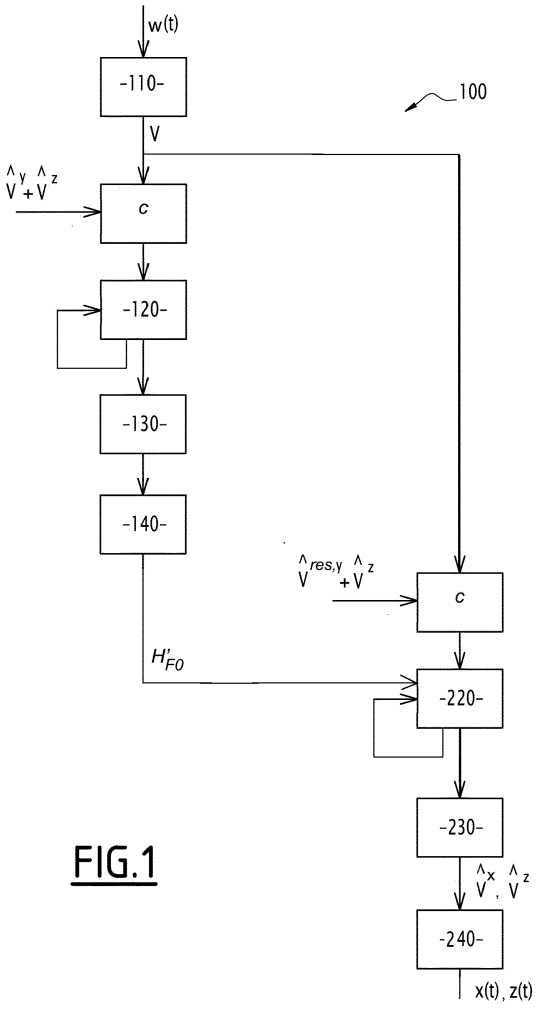

FIG.1

EP 3 040 989 A1

Description

[0001] La présente invention a pour domaine celui des procédés de séparation d'une pluralité de contributions dans un signal acoustique de mélange, et, en particulier, la séparation d'une contribution vocale, d'une contribution musicale de fond sonore, dans un signal acoustique de mélange.

[0002] Une bande son d'une chanson comporte une contribution vocale (les paroles chantées par un ou plusieurs chanteurs) et une contribution musicale (la musique d'accompagnement jouée par un ou plusieurs instruments).

[0003] Une bande son d'un film comporte une contribution vocale (les dialogues entre acteurs) superposée à une contribution musicale (les effets spéciaux sonores et/ou une musique de fond).

[0004] Il est connu des algorithmes de séparation permettant de séparer la contribution vocale, de la contribution musicale, dans une bande son originale.

[0005] Par exemple l'article de Jean-Louis Durrieu et al. "An iterative approach to monaural musical mixture de-soloing," in International Conference on Acoustics, Speech, and Signal Processing (ICASSP),Taipei, Taiwan, April 2009, pp. 105 - 108 divulgue un algorithme de séparation du type algorithme de séparation de sources sous-déterminée fondé sur une décomposition en matrices non-négatives, permettant de séparer la contribution vocale de la contribution de fond sonore.

[0006] Cependant, les algorithmes de séparation connus ne permettent pas de prendre correctement en compte le phénomène de réverbération affectant les composantes du mélange.

[0007] Dans le cas particulier d'une composante vocale, celle-ci résulte de la superposition de la voix sèche, ou pure dans ce qui suit, correspondant à l'enregistrement du son émis par le chanteur et qui s'est propagé directement vers le microphone d'enregistrement, et de la réverbération, correspondant à l'enregistrement du son émis par le chanteur mais qui s'est propagé indirectement vers le microphone d'enregistrement, c'est-à-dire par réflexion, éventuellement multiples, sur les parois de la salle d'enregistrement. La réverbération, constituée des échos de la voix pure à un instant donné, s'étale sur un intervalle de temps pouvant être important (par exemple trois secondes). Dit autrement, à un instant donné, la contribution vocale résulte de la superposition de la voix pure à cet instant et des différents échos de la voix pure à des instants précédents.

[0008] Or, les algorithmes de séparation connus ne prennent pas en compte les effets à long terme de la réverbération affectant une composante du mélange. Le document de Ngoc Q. K. Duong, Emmanuel Vincent, et Remi Gribonval, "Underdetermined reverberant audio source separation using a full-rank spatial covariance model," IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 7, pp. 1830 - 1840, Sept 2010 s'intéresse aux effets instantanés de diffusions spatiales de la réverbération, mais ne modélise pas les effets de mémoire, c'est-à-dire la prise en compte du temps de latence entre l'enregistrement d'un son et l'enregistrement des échos associé à ce son. Ainsi, le type d'algorithme proposé par ce document ne s'applique qu'à des signaux multicanaux et ne permet pas une extraction correcte des effets de réverbération, que l'on peut trouver dans la musique. Dans le cas d'une composante vocale, la réverbération qui affecte cette composante est répartie dans les différentes composantes obtenues à l'issue de la séparation. La composante vocale séparée perd de sa richesse et la composante musicale d'accompagnement n'est pas de bonne qualité.

[0009] Il est à noter que la réverbération peut avoir pour cause les conditions dans lesquelles est réalisée la prise de son, mais peut également être ajoutée artificiellement au cours de la post-production de la bande son, essentiellement pour des raisons esthétiques.

[0010] Il y a donc un besoin pour un procédé permettant de séparer des contributions dans un mélange, ces contributions intégrant une réverbération du signal sonore pure correspondant. Plus particulièrement, il y a un besoin pour séparer une contribution vocale pure affectée par de la réverbération, d'une contribution musicale de fond sonore, dans un signal sonore.

[0011] L'invention a donc pour but de pallier ce problème.

[0012] L'invention a donc pour objet un procédé de séparation et un produit programme conformes aux revendications.

[0013] L'invention sera mieux comprise à la lecture de la description qui va suivre d'un mode de réalisation particulier, donné uniquement à titre d'exemple illustratif et non limitatif, et faite en se référant aux dessins annexés sur lesquels :

- la figure 1 est une représentation sous forme de blocs des différentes étapes du procédé de séparation selon l'invention ; et,
- les figures 2 et 3 correspondent à des graphes qui résultent de tests permettant de comparer, selon des critères normatifs connus, les résultats de la mise en oeuvre du procédé de la figure 1.

[0014] En se référant à la figure 1, le procédé de séparation 100 utilise un signal acoustique temporel de mélange $w(t)$, pour délivrer un signal acoustique vocal $y(t)$ et un signal acoustique musical $z(t)$.

[0015] Les signaux sont tous des signaux acoustiques, de sorte que le qualificatif d'acoustique sera omis dans ce qui suit.

[0016] Ces signaux sont des signaux temporels. Ils dépendent du temps *t*.

[0017] Le signal acoustique de mélange est une bande son source, ou tout au moins un extrait d'une bande son.

[0018] Le signal acoustique de mélange *w*(*t*) comprend une première contribution dite spécifique et une seconde contribution dite d'accompagnement.

[0019] Dans la présente description, la première contribution est une contribution vocale et correspond à des paroles chantées par un chanteur.

[0020] La seconde contribution est une contribution musicale et correspond à l'accompagnement musical du chanteur.

[0021] Le signal acoustique vocal *y*(*t*) correspond à la seule contribution vocale, isolée du reste du signal de mélange *w*(*t*), et le signal acoustique musical *z*(*t*) correspond à la seule contribution musicale, isolée du reste du signal de mélange *w(t).*

[0022] Dans le présent mode de réalisation, on considère que seule la contribution vocale est réverbérée.

[0023] La réverbération est modélisée de la manière suivante :

$$y(t) = r(t) * x(t)$$

où *x(t)* est le signal vocal pur, c'est-à-dire le signal sonore généré par le chanteur est qui s'est propagé directement vers le microphone d'enregistrement ; et où *r*(*t*) est une réponse impulsionnelle, qui est une distribution donnant l'amplitude des échos pour chaque instant d'arrivé de l'écho correspondant sur le microphone d'enregistrement, et où * correspond au produit de convolution.

[0024] Le signal vocal pur *x(t)* est le signal en champ libre et la réponse impulsionnelle *r*(*t*) est caractéristique de l'environnement acoustique de l'enregistrement.

[0025] Dans le domaine temps fréquence, pour des spectrogrammes non-négatifs, ce modèle de réverbération peut être approximé, tel que proposé dans le document de Rita Singh, Bhiksha Raj, et Paris Smaragdis, "Latent-variable decomposition based dereverberation of monaural and multi-channel signals," in IEEE International Conference on Audio and Speech Signal Processing, Dallas, Texas, USA, March 2010, par :

$$V_{f,t}^{rev,y} = \sum_{\tau=1}^{T} V_{f,t-\tau+1}^{x} R_{f,\tau}$$

où $V^{rev,y}$ est le spectrogramme du signal y(t), considéré comme affecté par de la réverbération, V^{x} est le spectrogramme du signal x(t), *R* est une matrice *FxT* de réverbération correspondant au spectrogramme de la réponse impulsionnelle *r*(*t*), avec *F* la dimension fréquentiel et *T* la dimension temporelle de *R.*

[0026] La première étape 110 du procédé 100 consiste à échantillonner le signal de mélange *w*(*t*) et à calculer un spectrogramme *V* du signal de mélange *w(t).* De manière générale, un spectrogramme est défini comme la valeur absolue (ou bien le carré de la valeur absolue) de la transformée de Fourier à court terme d'un signal échantillonné. D'autres transformations temps-fréquence sont envisageables, telles qu'une transformée à Q constant, ou encore une transformée de Fourier à court terme suivie d'un filtrage fréquentiel (en utilisant un banc de filtres en échelle Mel ou Bark par exemple).

[0027] Pour chaque pas d'échantillonnage temporel, le spectrogramme comporte une trame en fréquence, indiquant pour chaque pas d'échantillonnage en fréquence, la puissance instantanée du signal.

[0028] Le spectrogramme *V* est donc une matrice F x U, de nombres réels positifs

[0029] U représente le nombre total de trames qui subdivisent la durée du signal du mélange *w(t).* F est le nombre total de pas d'échantillonnage en fréquence, qui vaut en général entre 200 et 2000.

[0030] Le procédé 100 comporte ensuite une première partie dans laquelle le signal vocal est considéré comme un signal vocal pur, sans réverbération.

[0031] Dans cette première partie, le spectrogramme de modélisation du signal de mélange est la somme du spectrogramme du signal vocal $\hat{V}^{y}$, et du spectrogramme du signal musical $\hat{V}^{z}$. $\hat{V}^{y}$ est le spectrogramme du signal y(t), considéré comme non affecté par de la réverbération. Cette modélisation est finalement la modélisation usuelle dans le cadre des méthodes de décomposition par factorisation en matrices non-négatives.

[0032] Il est à noter que *â* se réfère à une quantité qui est une estimation de la quantité *a.*

[0033] Ainsi, dans les étapes de la première partir du procédé 100, on cherche à estimer les deux spectrogrammes de sortie dont la somme approxime (signe ≈ dans l'expression suivante) au mieux le spectrogramme du mélange :

$$V \approx \hat{V} = \hat{V}^y + \hat{V}^z$$

[0034] La modélisation du signal vocal est fondée sur un modèle de production de la voix du type source / filtre, tel que proposé dans le document de Jean-Louis Durrieu et al. "An iterative approach to monaural musical mixture de-soloing," in International Conference on Acoustics, Speech, and Signal Processing (ICASSP),Taipei, Taiwan, April 2009, pp. 105 - 108 :

$$\hat{V}^y = (W_{F0}H_{F0})\odot(W_K H_K)$$

[0035] Le premier terme de cette modélisation est la source de la voix, qui correspond à l'excitation des cordes vocales : W_{F0} est une matrice d'atomes harmoniques, qui est prédéfinie et spécifique aux signaux vocaux ; H_{F0} est une matrice d'activation indiquant à chaque instant les atomes harmoniques de la matrice W_{F0} qui sont activés.

[0036] Le second terme de cette modélisation est le filtre de la voix, qui correspond au filtrage effectué par le conduit vocal : W_K est une matrice d'atomes de filtrage ; H_K est une matrice d'activation indiquant à chaque instant les atomes de filtrage de la matrice W_K qui sont activés.

[0037] L'opérateur $\odot$ correspond à la multiplication matricielle terme à terme de deux matrices (aussi dénommé produit d'Hadamard).

[0038] La modélisation du signal musical est fondée sur un modèle générique de factorisation par matrices non-négatives :

$$\hat{V}^z = (W_R H_R)$$

[0039] Des colonnes de W_R peuvent être vues comme des modèles spectraux élémentaires et H_R comme une matrice d'activation de ces modèles élémentaires au fil du temps.

[0040] La première partie du procédé consiste alors à estimer les matrices H_{F0}, W_K, H_K, W_R et H_R.

[0041] Afin d'estimer les paramètres de ces matrices, une fonction de coût C, fondée sur une divergence d par élément, est utilisée :

$$C = D(V|\hat{V}^y + \hat{V}^z) = \sum_{f,t} d(V_{ft}|\hat{V}^y_{ft} + \hat{V}^z_{ft})$$

[0042] Dans le mode de réalisation actuellement envisagé, la divergence d'Itakura-Saito, bien connue de l'homme du métier, est utilisée. Celle-ci est obtenue en fixant la valeur du paramètre de la beta-divergence à $\beta = 0$ et s'exprime donc :

$$d(a|b) = \frac{a}{b} - \log\frac{a}{b} - 1$$

[0043] Pour mémoire la beta-divergence est définie par :

$$d_\beta(a|b) = \begin{cases} \frac{1}{\beta(\beta-1)}\left(a^\beta + (\beta-1)b^\beta - \beta a b^{\beta-1}\right), & \beta \in \mathbb{R}\backslash\{0,1\} \\ a\log\frac{a}{b} - a + b, & \beta = 1 \\ \frac{a}{b} - \log\frac{a}{b} - 1, & \beta = 0 \end{cases}$$

où a et b sont deux scalaires réels positifs.

[0044] A l'étape 120, la fonction de coût C est ainsi minimisée de manière à déterminer la valeur optimale de chaque paramètre de chaque matrice. Cette minimisation est effectuée par itérations, avec des règles de mise à jour multiplicatives qui sont successivement appliquées à chacun des paramètres des matrices H_{F0}, W_K, H_K, W_R et H_R.

[0045] Ces règles de mise à jour sont par exemple élaborées en considérant le gradient (c'est-à-dire la dérivée partielle) de la fonction de coût C par rapport à chaque paramètre. Plus précisément, le gradient de la fonction de coût par rapport

au paramètre considéré est écrit sous la forme d'une différence entre deux termes positifs, et la règle de mise à jour correspondante est une multiplication du paramètre considéré par le rapport de ces deux termes.

[0046] Cela permet notamment que les paramètres restent non négatifs à chaque mise à jour et deviennent constants lorsque le gradient de la fonction de coût par rapport au paramètre considéré tend vers zéro.

[0047] De cette manière, les paramètres évoluent vers un minimum local.

[0048] Les règles de mise à jour sont ainsi les suivantes :

$$H_{F0} \leftarrow H_{F0} \odot \frac{W_{F0}^T\left((W_K H_K) \odot (V \odot \hat{V}^{\odot(\beta-2)})\right)}{W_{F0}^T\left((W_K H_K) \odot (\hat{V}^{\odot(\beta-1)})\right)}$$

$$H_K \leftarrow H_K \odot \frac{W_K^T\left((W_{F0} H_{F0}) \odot (V \odot \hat{V}^{\odot(\beta-2)})\right)}{W_K^T\left((W_{F0} H_{F0}) \odot (\hat{V}^{\odot(\beta-1)})\right)}$$

$$W_K \leftarrow W_K \odot \frac{\left((W_{F0} H_{F0}) \odot (V \odot \hat{V}^{\odot(\beta-2)})\right) H_K^T}{\left((W_{F0} H_{F0}) \odot (\hat{V}^{\odot(\beta-1)})\right) H_K^T}$$

$$H_R \leftarrow H_R \odot \frac{W_R^T\left(V \odot \hat{V}^{\odot(\beta-2)}\right)}{W_R^T\left(\hat{V}^{\odot(\beta-1)}\right)}$$

$$W_R \leftarrow W_R \odot \frac{\left(V \odot \hat{V}^{\odot(\beta-2)}\right) H_R^T}{\left(\hat{V}^{\odot(\beta-1)}\right) H_R^T}$$

où $\odot$ est un opérateur correspondant au produit terme à terme entre matrices (ou vecteur) ; $.\odot(.)$ est un opérateur correspondant à l'exponentiation terme à terme d'une matrice par un scalaire ; $(.)^T$ est la transposée d'une matrice.

[0049] Pour cette première partie, tous les paramètres sont initialisés avec des valeurs non-négatives choisies de manière aléatoire.

[0050] Puis, à l'étape 130, la matrice H_{F0} est contrainte en utilisant un algorithme de suivi tel que l'algorithme de « tracking » de Viterbi afin de sélectionner, pour chaque pas temporel, le pas en fréquence dans lequel on retrouve un maximum de puissance, sans être trop éloigné en fréquence des maxima de puissance sélectionnés pour les pas temporels précédents.

[0051] Puis, à l'étape 140, les coefficients de la matrice H_{F0} qui sont à une distance en fréquence supérieure à une distance de référence sont fixés à 0.

[0052] Une matrice H'_{F0} est obtenue.

[0053] Dans la seconde partie du procédé 100, le signal vocal est considéré comme affecté par de la réverbération. Il est à noter que la première partie du procédé permet d'obtenir des valeurs initiales pour les paramètres qui vont être estimés par itérations successives lors de la mise en oeuvre de la seconde partie du procédé. D'autres manières de définir les valeurs initiales de ces paramètres sont envisageables.

[0054] Dans cette seconde partie, la modélisation du signal vocal considéré comme réverbéré, $\hat{V}^{rev,y}$, en fonction du signal vocal pure $\hat{V}^x$ s'écrit alors :

$$[\hat{V}^{rev,y}]_{f,t} = [\hat{V}^x *_t R\ ]_{f,t} = \sum_{\tau=1}^{T} \hat{V}^x_{f,t-\tau+1} R_{f,\tau}$$

où $*_t$ dénote un opérateur de convolution ligne par ligne tel qu'explicité dans le membre de droite de l'équation ci-dessus.

[0055] La matrice de réverbération R comporte T pas de temps (de même durée qu'un pas d'échantillonnage du signal

de mélange), et *F* pas d'échantillonnage en fréquence. *T* est prédéterminé par l'utilisateur et vaut généralement entre 20 et 200, par exemple 100.

[0056] De plus, comme ci-dessus, le spectrogramme $\hat{V}^x$ du signal pure est modélisé par :

$$\hat{V}^x = (W_{F0}H_{F0}) \odot (W_K H_K)$$

[0057] La seconde partie du procédé consiste alors à estimer les matrices H_{F0}, W_K, H_K, W_R, H_R et R qui permettent d'approximer le spectrogramme du mélange V :

$$V \approx \hat{V}^{rev} = \hat{V}^{rev,y} + \hat{V}^z$$

[0058] Afin d'estimer les paramètres de ces matrices, une fonction de coût *C,* fondée sur une divergence d par élément, est utilisée :

$$C = D(V|\hat{V}^{rev,y} + \hat{V}^z) = \sum_{f,t} d(V_{ft}|\hat{V}_{ft}^{rev,y} + \hat{V}_{ft}^z)$$

[0059] Dans le mode de réalisation actuellement envisagé, la divergence d'Itakura-Saito, bien connue de l'homme du métier, est utilisée. Celle-ci est obtenue en fixant la valeur du paramètre de la beta-divergence à $\beta = 0$ et s'exprime donc :

$$d(a|b) = \frac{a}{b} - \log\frac{a}{b} - 1$$

[0060] Avantageusement, la fonction de coût de la seconde partie est similaire à celle utilisée dans la première partie.

[0061] A l'étape 220, la fonction de coût C est alors minimisée de manière à déterminer la valeur optimale de chaque paramètre de chaque matrice, en particulier les paramètres de la matrice de réverbération.

[0062] Cette minimisation est effectuée par itérations avec des règles de mise à jour multiplicatives, qui sont successivement appliquées à chacun des paramètres des matrices. Pour les matrices de la composante vocale intégrant une réverbération, on a :

$$R \leftarrow R \odot \frac{\left(V \odot \hat{V}^{rev\odot(\beta-2)}\right) *_t \hat{V}^x}{\hat{V}^{rev\odot(\beta-1)} *_t \hat{V}^x}$$

$$H_{F0} \leftarrow H_{F0} \odot \frac{W_{F0}^T\left((W_K H_K) \odot (R *_t \left(V \odot \hat{V}^{rev\odot(\beta-2)}\right))\right)}{W_{F0}^T\left((W_K H_K) \odot (R *_t \hat{V}^{rev\odot(\beta-1)})\right)}$$

$$H_K \leftarrow H_K \odot \frac{W_K^T\left((W_{F0} H_{F0}) \odot (R *_t \left(V \odot \hat{V}^{rev\odot(\beta-2)}\right))\right)}{W_K^T\left((W_{F0} H_{F0}) \odot (R *_t \hat{V}^{rev\odot(\beta-1)})\right)}$$

$$W_K \leftarrow W_K \odot \frac{\left((W_{F0} H_{F0}) \odot (R *_t \left(V \odot \hat{V}^{rev\odot(\beta-2)}\right))\right) H_K^T}{\left((W_{F0} H_{F0}) \odot (R *_t \hat{V}^{rev\odot(\beta-1)})\right) H_K^T}$$

où $*_t$ désigne l'opérateur de convolution ligne par ligne tel que défini ci-dessus.

[0063] Pour la composante musicale de fond sonore, on a, comme dans la première partie du procédé :

$$H_R \leftarrow H_R \odot \frac{W_R^T\left(V \odot \hat{V}^{rev\odot(\beta-2)}\right)}{W_R^T\left(\hat{V}^{rev\odot(\beta-1)}\right)}$$

$$W_R \leftarrow W_R \odot \frac{\left(V \odot \hat{V}^{rev\odot(\beta-2)}\right)H_R^T}{\left(\hat{V}^{rev\odot(\beta-1)}\right)H_R^T}$$

[0064] Comme indiqué ci-dessus, les règles de mise à jour sont élaborées à partir de la dérivée partielle de la fonction de coût C par rapport à chaque paramètre pertinent. Elles dépendent donc de la forme retenue pour la fonction de coût, notamment la divergence utilisée dans cette fonction de coût. Les règles ci-dessus sont donc spécifiques de l'utilisation d'une béta-divergence.

[0065] Il est à noter que, puisque ces règles résultent chacune d'une dérivation partielle selon un paramètre spécifique, la règle de mise à jour de la matrice de réverbération R est générale au sens où elle ne dépend pas de la modélisation retenue pour le spectrogramme $\hat{V}^x$ du signal pure ou celle du spectrogramme $\hat{V}^z$ de fond sonore.

[0066] En ce qui concerne la matrice H_{F0}, les itérations partent de la matrice H'_{F0} déterminée dans la première partie du procédé. Il est à noter que, puisque les règles de mise à jour sont multiplicatives, les coefficients de la matrice H_{F0} fixés initialement à 0 resteront à 0 au cours de la minimisation de la fonction de coût dans la seconde partie du procédé.

[0067] Les autres paramètres de la modélisation et, en particulier ceux du spectrogramme de la contribution spécifique réverbérée $\hat{V}^{rev,y}$ sont initialisés avec des valeurs aléatoires.

[0068] Lorsque la distance entre le spectrogramme de mélange V et le spectrogramme estimé $\hat{V} = \hat{V}^{rev,y} + \hat{V}^z$ est inférieure à un seuil prédéterminé ou lorsqu'un nombre d'itérations limite fixé à l'avance est atteint, le procédé sort de la boucle d'itération et les valeurs des matrices obtenues, R, H_{F0}, W_K, H_K, W_R et H_R, sont les valeurs finales.

[0069] A l'étape 230, des traitements adaptés classiques (en particulier un traitement du type filtrage de Wiener) sont appliqués sur les spectrogrammes précédents pour obtenir notamment les spectrogrammes d'intérêt $\hat{V}^x$, $\hat{V}^z$. Puis, à l'étape 240, une transformation inverse de celle de l'étape 110 est réalisée sur ces spectrogrammes pour obtenir les signaux de sorties, signal vocal pur x(t) et signal musical z(t).

[0070] Dans les modes de réalisation décrits ici en détail, ces signaux acoustiques sont des signaux monophoniques. En variante, ces signaux sont stéréophoniques. Plus généralement encore, ils sont multicanaux. L'homme du métier sait comment adapter à des signaux stéréophoniques ou multicanaux les traitements présentés pour le cas de signaux monophoniques.

[0071] Le mode de réalisation préféré est relatif à une composante spécifique ou d'intérêt qui est une composante vocale. Cependant, la modélisation de la réverbération d'une composante est générale et s'applique à tout type de composante. En particulier, la composante de fond sonore peut également être affectée par une réverbération.

[0072] De plus, n'importe quel type de modélisations non-négatives des spectrogrammes des sons non réverbérés peut également être utilisées, en lieu et place de celles utilisées ci-dessus.

[0073] Par ailleurs, dans le mode de réalisation présenté ci-dessus, le mélange comporte deux composantes. La généralisation à un nombre quelconque de composantes est directe.

[0074] Des tests comparatifs ont été menés afin de comparer les résultats de la mise en oeuvre du présent procédé :

- le premier procédé est une séparation, fondée sur une méthode de type NMF, sans inclure de modélisation sur la réverbération ;
- le second procédé est une séparation selon le procédé décrit ci-dessus, c'est-à-dire incluant une modélisation de la réverbération du signal vocal ; et,
- le troisième procédé est une limité mathématique théorique.

[0075] Afin de quantifier les résultats obtenus pour les différents procédés, des indicateurs standards du domaine de la séparation de sources ont été calculés. Ces indicateurs sont le rapport signal sur distorsion SDR (selon l'acronyme anglais « Signal to Distorsion Ratio »), et qui correspond à un test quantitatif ; le rapport signal sur artefact SAR (selon l'acronyme « Signal to Artefact Ratio »), et qui correspond aux artefacts dans les composantes séparées ; et le rapport signal sur interférence SIR (selon l'acronyme anglais « Signal to Interference Ratio »), et qui correspond aux interférences résiduelles entre les composantes séparées.

[0076] Les résultats sont présentés sur les figures 2 pour le signal vocal et la figure 3 pour le signal musical.

[0077] Le procédé selon l'invention améliore donc les résultats obtenus, quelle que soit la manière de les analyser.

Revendications

1. Procédé de séparation (100), dans un signal acoustique de mélange w(t), d'une contribution spécifique pure, affectée par de la réverbération, et d'une contribution de fond sonore,
caractérisé en ce qu'il consiste à séparer la contribution spécifique pure *x(t)* et la contribution de fond sonore *z*(*t*), en utilisant un spectrogramme de modélisation du signal acoustique de mélange $\hat{V}^{rev}$ correspondant à la somme d'un spectrogramme d'une contribution spécifique réverbérée $\hat{V}^{rev,y}$ et d'un spectrogramme de la contribution de fond sonore $\hat{V}^{z}$, le spectrogramme de la contribution spécifique réverbérée dépendant du spectrogramme de la contribution spécifique pure $\hat{V}^{x}$ selon le modèle :

$$\hat{V}_{f,t}^{rev,y} = \sum_{\tau=1}^{T} \hat{V}_{f,t-\tau+1}^{x} R_{f,\tau}$$

où *R* est une matrice *FxT* de réverbération, *f* est un indice de fréquence, *t* est un indice de temps, et τ un entier entre 1 et *T* ; et
en calculant de manière itérative une estimation du spectrogramme de la contribution de fond sonore $\hat{V}^{z}$, du spectrogramme de la contribution spécifique pure $\hat{V}^{x}$ et de la matrice de réverbération *R* de manière à minimiser une fonction de coût (*C*) entre un spectrogramme du signal de mélange *V* et le spectrogramme de modélisation du signal de mélange $\hat{V}^{rev}$.

2. Procédé selon la revendication 1, **caractérisé en ce que** la fonction de coût (*C*) utilise une divergence (*d*) entre le spectrogramme du signal de mélange et le spectrogramme de modélisation du signal de mélange, notamment la divergence dite beta-divergence définie par :

$$d_\beta(a|b) = \begin{cases} \frac{1}{\beta(\beta-1)}\left(a^\beta + (\beta-1)b^\beta - \beta a b^{\beta-1}\right), & \beta \in \mathbb{R}\backslash\{0,1\} \\ a\log\frac{a}{b} - a + b, & \beta = 1 \\ \frac{a}{b} - \log\frac{a}{b} - 1, & \beta = 0 \end{cases}$$

où *a* et *b* sont deux scalaires réels positifs.

3. Procédé selon la revendication 2, **caractérisé en ce que** la minimisation de la fonction de coût met en oeuvre, pour obtenir une estimation de la matrice de réverbération, des règles de mise à jour multiplicatives du type :

$$R \leftarrow R \odot \frac{\left(V \odot \hat{V}^{rev\odot(\beta-2)}\right) *_t \hat{V}^{x}}{\hat{V}^{rev\odot(\beta-1)} *_t \hat{V}^{x}}$$

Avec $\hat{V}^{rev} = \hat{V}^{rev,y} + \hat{V}^{z}$; et où $\odot$ est un opérateur correspondant au produit terme à terme entre matrices (ou vecteur) ; $.\odot(.)$ est un opérateur correspondant à l'exponentiation terme à terme d'une matrice par un scalaire ; $*_t$ est un opérateur de convolution temporelle entre deux matrices défini par $[A *_t B]_{f,\tau} = \sum_{\tau=t}^{T} A_{f,\tau} B_{f,\tau-t+1}$.

4. Procédé selon l'une quelconque des revendications précédentes, **caractérisé en ce que**, la contribution spécifique pure étant une contribution vocale, le spectrogramme de la contribution spécifique pure $\hat{V}^{x}$ est modélisé par :

$$\hat{V}^x = (W_{F0}H_{F0}) \odot (W_K H_K)$$

où W_{F0} est une matrice d'atomes harmoniques prédéfinie, H_{F0} est une matrice d'activation des atomes harmoniques de la matrice W_{F0}, W_K est une matrice d'atomes de filtrage, H_K est une matrice d'activation des atomes de filtrage de la matrice W_K, et où $\odot$ est un opérateur correspondant au produit terme à terme entre matrices.

5. Procédé selon la revendication 3 et la revendication 4, **caractérisé en ce que** la minimisation de la fonction de coût met en oeuvre des règles de mise à jour multiplicatives du type :

$$H_{F0} \leftarrow H_{F0} \odot \frac{W_{F0}^T\left((W_K H_K) \odot \left(R *_t \left(V \odot \hat{V}^{rev\odot(\beta-2)}\right)\right)\right)}{W_{F0}^T\left((W_K H_K) \odot \left(R *_t \hat{V}^{rev\odot(\beta-1)}\right)\right)}$$

$$H_K \leftarrow H_K \odot \frac{W_K^T\left((W_{F0} H_{F0}) \odot \left(R *_t \left(V \odot \hat{V}^{rev\odot(\beta-2)}\right)\right)\right)}{W_K^T\left((W_{F0} H_{F0}) \odot \left(R *_t \hat{V}^{rev\odot(\beta-1)}\right)\right)}$$

$$W_K \leftarrow W_K \odot \frac{\left((W_{F0} H_{F0}) \odot \left(R *_t \left(V \odot \hat{V}^{rev\odot(\beta-2)}\right)\right)\right) H_K^T}{\left((W_{F0} H_{F0}) \odot \left(R *_t \hat{V}^{rev\odot(\beta-1)}\right)\right) H_K^T}$$

Avec $\hat{V}^{rev} = \hat{V}^{rev'y} + \hat{V}^z$; et où $\odot$ est un opérateur correspondant au produit terme à terme entre matrices (ou vecteur) ; $.\odot(.)$ est un opérateur correspondant à l'exponentiation terme à terme d'une matrice par un scalaire ; $(.)^T$ est la transposée d'une matrice ; $*_t$ est un opérateur de convolution temporelle entre deux matrices défini par $[A *_t B]_{f,\tau} = \sum_{\tau=t}^{T} A_{f,\tau} B_{f,\tau-t+1}$.

6. Procédé selon l'une quelconque des revendications précédentes, **caractérisé en ce que** le spectrogramme de la contribution de fond sonore $\hat{V}^z$ est modélisé par une factorisation en matrices non-négatives :

$$\hat{V}^z = (W_R H_R)$$

où W_R est une matrice de modèles spectraux élémentaires et H_R est une matrice d'activation des modèles spectraux élémentaires de la matrice W_R.

7. Procédé selon la revendication 2 et la revendication 6, **caractérisé en ce que** la minimisation de la fonction de coût met en oeuvre des règles de mise à jour multiplicatives du type :

$$H_R \leftarrow H_R \odot \frac{W_R^T\left(V \odot \hat{V}^{rev\odot(\beta-2)}\right)}{W_R^T\left(\hat{V}^{rev\odot(\beta-1)}\right)}$$

$$W_R \leftarrow W_R \odot \frac{(V \odot \hat{V}^{rev\odot(\beta-2)}) H_R^T}{\left(\hat{V}^{rev\odot(\beta-1)}\right) H_R^T}$$

Avec $\hat{V}^{rev} = \hat{V}^{rev,y} + \hat{V}^{Z}$; et où $\odot$ est un opérateur correspondant au produit terme à terme entre matrices (ou vecteur) ; $.\odot(.)$ est un opérateur correspondant à l'exponentiation terme à terme d'une matrice par un scalaire ; $(.)^T$ est la transposée d'une matrice.

8. Procédé selon l'une quelconque des revendications précédentes, **caractérisé en ce que**, la séparation de la contribution spécifique pure *x(t)* et la contribution de fond sonore *z*(*t*) en utilisant un spectrogramme de modélisation du signal acoustique de mélange $\hat{V}^{rev}$ constituant une seconde partie du procédé, celui-ci comporte une première partie consistant à séparer, dans le signal acoustique de mélange w(t), une contribution spécifique et une contribution du fond sonore, sans tenir compte de la réverbération, des paramètres d'initialisation parmi les paramètres obtenus à l'issue de la première partie du procédé étant utilisés comme valeur initiale des paramètres correspondants dans le spectrogramme de la contribution spécifique réverbérée $\hat{V}^{rev,y}$ de la seconde partie du procédé.

9. Procédé selon la revendication 8, **caractérisé en ce que** la première partie comporte la minimisation d'une fonction de coût mettant en oeuvre un algorithme similaire à celui mis en oeuvre dans la seconde partie.

10. Procédé selon la revendication 2 et la revendication 9, **caractérisé en ce que**, pour la minimisation de la fonction de coût, la première partie du procédé met en oeuvre des règles de mise à jour multiplicatives du type :

$$H_{F0} \leftarrow H_{F0} \odot \frac{W_{F0}^T\left((W_K H_K) \odot (V \odot \hat{V}^{\odot(\beta-2)})\right)}{W_{F0}^T\left((W_K H_K) \odot (\hat{V}^{\odot(\beta-1)})\right)}$$

$$H_K \leftarrow H_K \odot \frac{W_K^T\left((W_{F0} H_{F0}) \odot (V \odot \hat{V}^{\odot(\beta-2)})\right)}{W_K^T\left((W_{F0} H_{F0}) \odot (\hat{V}^{\odot(\beta-1)})\right)}$$

$$W_K \leftarrow W_K \odot \frac{\left((W_{F0} H_{F0}) \odot (V \odot \hat{V}^{\odot(\beta-2)})\right) H_K^T}{\left((W_{F0} H_{F0}) \odot (\hat{V}^{\odot(\beta-1)})\right) H_K^T}$$

$$H_R \leftarrow H_R \odot \frac{W_R^T\left(V \odot \hat{V}^{\odot(\beta-2)}\right)}{W_R^T\left(\hat{V}^{\odot(\beta-1)}\right)}$$

$$W_R \leftarrow W_R \odot \frac{\left(V \odot \hat{V}^{\odot(\beta-2)}\right) H_R^T}{\left(\hat{V}^{\odot(\beta-1)}\right) H_R^T}$$

avec : $\hat{V} = \hat{V}^{x} + \hat{V}^{Z}$, $\hat{V}^{z} = (W_R H_R)$, et $\hat{V}^{x} = (W_{F0} H_{F0}) \odot (W_K H_K)$; où W_R est une matrice de modèles spectraux élémentaires et H_R est une matrice d'activation des modèles spectraux élémentaires de la matrice W_R ; où W_{F0} est une matrice d'atomes harmoniques prédéfinie, H_{F0} est une matrice d'activation des atomes harmoniques de la matrice W_{F0}, W_K est une matrice d'atomes de filtrage, H_K est une matrice d'activation des atomes de filtrage de la matrice W_K ; et où $\odot$ est un opérateur correspondant au produit terme à terme entre matrices (ou vecteur) ; $.\odot(.)$ est un opérateur correspondant à l'exponentiation terme à terme d'une matrice par un scalaire ; $(.)^T$ est la transposée d'une matrice.

11. Procédé selon l'une quelconque des revendications 8 à 10, **caractérisé en ce qu'**il comporte, dans la première partie du procédé, à la suite de la minimisation de la fonction de coût, l'application d'un algorithme de suivi du maximum de puissance dans la matrice d'activation de la contribution spécifique H_{F0}, ledit algorithme étant de préférence du type algorithme de Viterbi, puis la mise à zéro de tous les termes de la matrice d'activation de la

contribution spécifique H_{F0} qui sont trop éloignés du maximum de puissance trouvé, les termes de la matrice d'activation de la contribution spécifique H_{F0} constituant les paramètres d'initialisation qui sont utilisés comme valeur initiale des paramètres correspondants dans le spectrogramme de la contribution spécifique réverbérée $\hat{V}^{rev,y}$ de la seconde partie du procédé, les autres paramètres du spectrogramme de la contribution spécifique réverbérée $\hat{V}^{rev,y}$ étant initialisés avec des valeurs aléatoires.

12. Produit programme d'ordinateur, **caractérisé en ce qu'**il comporte des instructions propres à être stockées dans la mémoire d'un calculateur et exécutées par le processeur dudit calculateur pour mettre en oeuvre un procédé de séparation conforme à l'une quelconque des revendications précédentes.

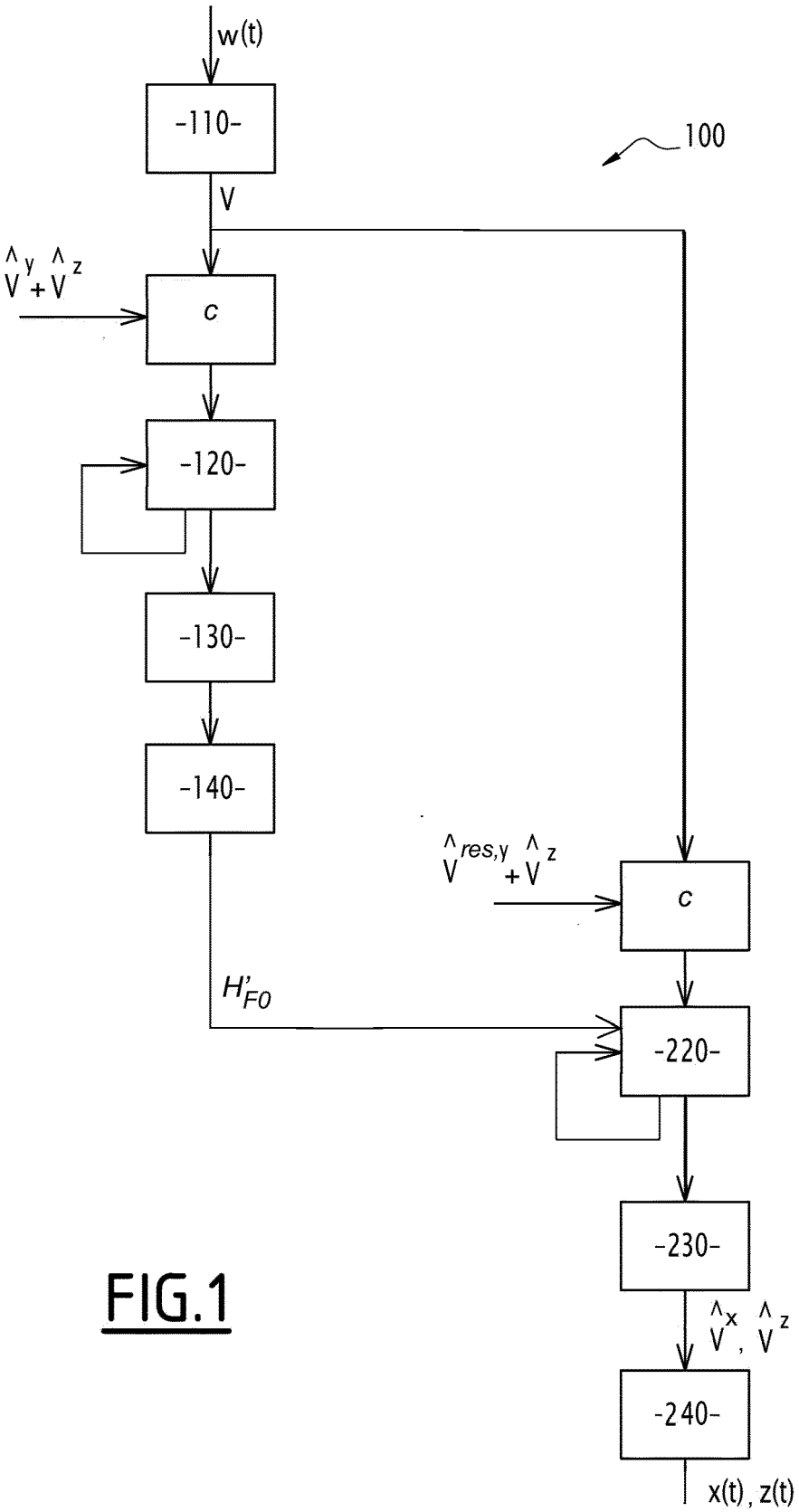
w(t)
-110-
100
V
$\hat{V}_y + \hat{V}_z$
c
-120-
-130-
-140-
$\hat{V}_{res,y} + \hat{V}_z$
c
H'_{F0}
-220-
-230-
$\hat{V}_x, \hat{V}_z$
-240-
x(t), z(t)

FIG.1

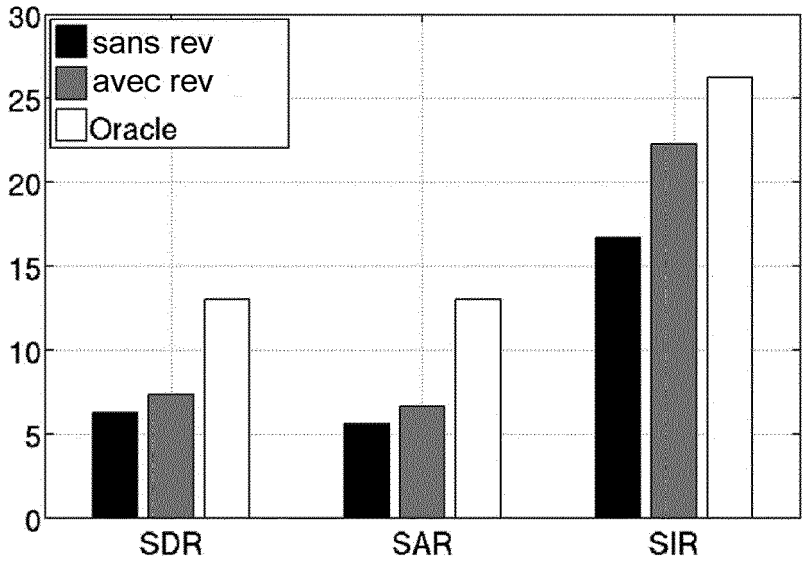
30
25
20
15
10
5
0
sans rev
avec rev
Oracle
SDR
SAR
SIR

FIG.2

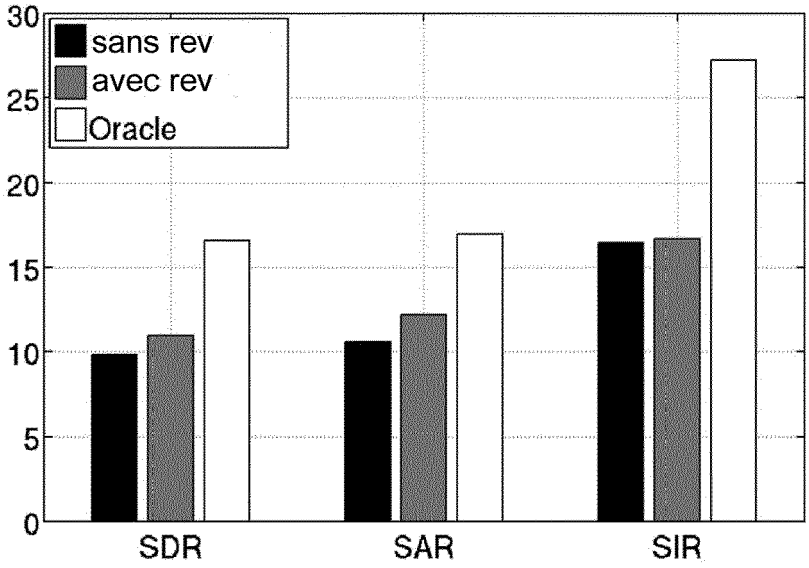
30
25
20
15
10
5
0
sans rev
avec rev
Oracle
SDR
SAR
SIR

FIG.3

Europäisches Patentamt
European Patent Office
Office européen des brevets

RAPPORT DE RECHERCHE EUROPEENNE

Numéro de la demande
EP 15 19 8713

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,D	JEAN-LOUIS DURRIEU ET AL: "An iterative approach to monaural musical mixture de-soloing", ACOUSTICS, SPEECH AND SIGNAL PROCESSING, 2009. ICASSP 2009. IEEE INTERNATIONAL CONFERENCE ON, IEEE, PISCATAWAY, NJ, USA, 19 avril 2009 (2009-04-19), pages 105-108, XP031459177, ISBN: 978-1-4244-2353-8	1,2,4,6, 12	INV. G10L21/0208
A	* Section 2 *	3,5,7-11	

A	RITA SINGH ET AL: "Latent-variable decomposition based dereverberation of monaural and multi-channel signals", ACOUSTICS SPEECH AND SIGNAL PROCESSING (ICASSP), 2010 IEEE INTERNATIONAL CONFERENCE ON, IEEE, PISCATAWAY, NJ, USA, 14 mars 2010 (2010-03-14), pages 1914-1917, XP031697281, ISBN: 978-1-4244-4295-9 * page 1915 *	1,12	
	-----		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
La Haye	3 mai 2016	Taddei, Hervé

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)

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.

Littérature non-brevet citée dans la description

- **JEAN-LOUIS DURRIEU et al.** An iterative approach to monaural musical mixture de-soloing. *International Conference on Acoustics, Speech, and Signal Processing (ICASSP),Taipei, Taiwan,* Avril 2009, 105-108 **[0005] [0034]**
- **NGOC Q. K. DUONG ; EMMANUEL VINCENT ; REMI GRIBONVAL.** Underdetermined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Transactions on Audio, Speech, and Language Processing,* Septembre 2010, vol. 18 (7), 1830-1840 **[0008]**
- **RITA SINGH ; BHIKSHA RAJ ; PARIS SMARAGDIS.** Latent-variable decomposition based dereverberation of monaural and multi-channel signals. *IEEE International Conference on Audio and Speech Signal Processing, Dallas, Texas, USA,* Mars 2010 **[0025]**