Détection et classification des sons : application aux sons de la vie courante et à la parole

Dan ISTRATE, Michel VACHER, Jean François SERIGNAT

Laboratoire CLIPS-IMAG, UMR CNRS-UJF-INPG 5524
385, rue de la Bibliothèque
B.P. 53 - 38041 Grenoble Cedex 9
Dan.Istrate@imag.fr,Michel.Vacher@imag.fr
Jean-Francois.Serignat@imag.fr

Résumé – Depuis quelques années se développe le concept général d'espace perceptif (salle intelligente) qui répond de diverses façons aux besoins, demandes, attentes des acteurs humains. Un système d'extraction de l'information du son à trois étapes est proposé. La première étape, permet la détection et l'extraction des sons du flux sonore continu. L'algorithme de détection proposé est basé sur la transformée en ondelettes, il permet de s'affranchir du bruit et d'obtenir une bonne résolution temporelle. La deuxième étape utilise un mélange de distributions de Gauss (GMM) pour faire la classification du signal sonore entre parole et sons et aiguiller le signal sur le processus adapté : reconnaissance de la parole (non traitée dans l'article) ou classification des sons. La troisième étape, celle de classification des sons de la vie courante, est aussi réalisée avec un système à base de GMM. Les paramètres acoustiques sont étudiés étant donné qu'ils ont une influence essentielle sur le système de classification ; par ailleurs, de nouveaux paramètres issus de la transformée en ondelettes sont proposés. Chaque étape de l'étude est validée au moyen d'un corpus spécifique.

Abstract – Recently, the general concept of perceptive spaces or smart rooms is in a continuous development and tries to answer in different ways to the needs, demands or expectations of human actors. This paper presents a system to extract information from sound signals, which contains three stages. The first stage, sound event detection, takes care of the sound detection and extraction from a continuous acoustic flux and it uses an algorithm based on the wavelet transform. The algorithm described in the paper offers good temporal resolution and performances in a noisy environment. The second stage aims to Speech/Sound classification and uses Gaussian Mixtures Models (GMM). The third stage is also based on GMM and realizes the classification of the everyday life sounds. The acoustical parameters are studied since they have an important influence on the classification performances. New parameters based on the wavelet transform are proposed. Every stage of the system is validated on a specific corpus.

1 Introduction

Depuis quelques années se développe le concept général d'espace perceptif (salle intelligente) qui répond de diverses façons aux besoins, demandes, attentes des acteurs humains. Les espaces perceptifs exploitent des signaux de parole, des signaux vidéo, des données de l'environnement afin de permettre la localisation des personnes, la reconnaissance des gestes, etc.

L'analyse et l'extraction des informations du son peuvent être très utiles dans ces espaces. La reconnaissance de la parole et du locuteur sont fréquemment étudiées. La possibilité d'identification des sons de la vie courante comme les claquements de porte, les chutes d'objets, les sons de vaisselle est utile dans le cadre des applications de télésurveillance médicale ou pour l'indexation des bases de données, mais il s'agit d'un domaine encore peu exploré. Cet article propose un système d'extraction d'informations du son acquis par un système multicanal avec pour principale application la télésurveillance médicale.

Le système d'analyse sonore proposé est présenté sur la figure 1, il se décompose en quatre modules pour permettre le fonctionnement du système de traitement en temps réel. Le premier module, ou module de détection, a pour rôle d'extraire les signaux à identifier (sons de la vie courante ou parole) du bruit environnemental. Le deuxième module doit déterminer si le signal extrait appartient à la classe des sons de la vie courante ou à celle de la parole. Suivant ce résultat, le module utilisé sera soit un module de classification, soit un système de reconnaissance de la parole continue qui n'est pas décrit ici.

2 Détection des évènements sonores

Le rôle de la détection est de déterminer l'instant d'apparition d'un événement sonore, en vue de l'extraire du bruit de fond pour un traitement ultérieur de classification. Nous considérons comme événement sonore les sons de la vie courante ou la parole noyés dans un bruit non-stationnaire.

L'adaptation du seuil à chaque analyse en utilisant la probabilité de distribution de l'amplitude du signal de parole est une des méthodes classiques de détection. Il existe aussi des méthodes de détection de la voix (VAD-Voice Activity Detection) mais celles-ci utilisent des propriétés spécifiques à la parole, comme par exemple, la présence de la fréquence fondamentale ou les statistiques d'ordres supérieurs sur le résidu de la prédiction linéaire pour différencier la parole du bruit. Ces méthodes se basent soit sur le modèle de production de la parole, soit sur d'autres propriétés du signal de parole et ne sont pas utilisables

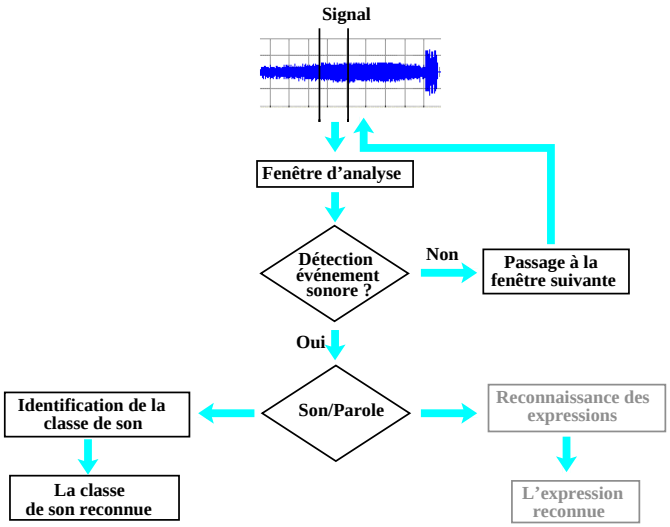

FIG. 1 – Le système d'analyse sonore proposé

pour les sons de la vie courante.

Les travaux qui se rapprochent le plus de notre problème ont été conduits par A. Dufaux qui a étudié la détection et la classification de sons impulsionnels pour un système anti-effraction [1]. Les techniques de détection des signaux impulsionnels proposés utilisaient l'énergie du signal et faisaient ressortir les maxima locaux par un traitement adéquat (filtrage médian, calcul des paramètres statistiques, etc.). Les résultats dépendent beaucoup des propriétés du bruit de fond.

L'algorithme de détection développé pour notre application devra fonctionner en présence d'un bruit environnemental basse fréquence et non stationnaire important pour lequel le système cité a des performances insuffisantes, en effet le rapport signal sur bruit peut descendre jusqu'à 0 dB.

Algorithme proposé. La détection du signal utile est facilitée par une décomposition en fréquence du flux sonore. Dans cette étude, la transformée en ondelettes discrète (DWT) est préférée à la transformée de Fourier discrète (FFT) car la résolution temporelle en hautes fréquences est une fraction de la largeur de la fenêtre d'analyse ; en effet, une détection précise du début et de la fin du signal est importante pour les étapes qui suivent.

En effet, le pavage temps-fréquence n'est pas uniforme pour la transformée en ondelettes. La résolution temporelle est alors supérieure pour les hautes fréquences, à l'opposé de la résolution fréquentielle, car le nombre d'ondelettes est d'autant plus important que le support de l'ondelette se situe haut en fréquence [2]. Les ondelettes de Daubechies sont choisies en traitement du signal à cause de leurs propriétés spectrales ; l'algorithme proposé utilise les ondelettes de Daubechies avec 6 moments nuls.

L'algorithme proposé atténue l'influence du bruit environnemental en utilisant seulement les ondelettes des trois coefficients de plus hautes fréquences de la transformée. La taille de la fenêtre d'analyse est imposée par l'application temps réel, 128 ms soit 2048 échantillons pour une fréquence d'échantillonnage de 16 kHz ; la transformée en ondelettes permet un décalage temporel à l'intérieur de cette fenêtre qui rend possible une détermination plus fine de l'instant d'apparition du son. L'algorithme, comme le montre la figure 2, commence par

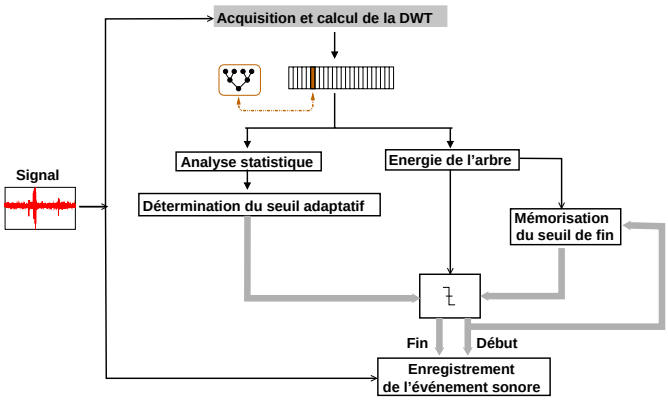

FIG. 2 – Le schéma de l'algorithme de détection basé sur la transformée en ondelettes

le calcul de la transformée en ondelettes sur la fenêtre d'analyse, puis le calcul de l'énergie des arbres d'ondelettes avec une profondeur de 3 pour des trames ayant une largeur de 32 ms et se chevauchant de 16 ms. La détection est réalisée par l'application d'un seuil adaptatif sur la valeur de l'énergie des arbres d'ondelettes de la trame. Le seuil adaptatif dépend de la moyenne des valeurs de l'énergie précédentes et d'une constante.

Si le signal extrait comprend des parties composées seulement de bruit, ceci aura une influence sur la classification et fera baisser les performances ; si une faible partie du signal est coupée, les effets seront moins importants car l'évolution temporelle du signal n'est pas prise en compte par le système de classification utilisé. La détection de fin de signal est obtenue par un deuxième seuil sur l'énergie de l'arbre d'ondelettes. Ce seuil est obtenu à partir de la valeur de l'énergie de l'arbre au moment de la détection du début de l'événement sonore. Pour tenir compte des périodes de silence dans le signal de parole, la fin du signal est considérée atteinte seulement si l'énergie reste en-dessous du seuil de fin pendant un nombre suffisant de trames consécutives (12 trames ≈192ms) ; on considère comme la fin du signal l'instant de la première valeur en-dessous du seuil.

Un exemple de détection d'un événement sonore (l'expression « Appelez quelqu'un ») mélangé avec le bruit de l'appartement de test (que nous appelerons à la suite HIS) est présenté dans la figure 3. Le rapport signal sur bruit (RSB) est de 0 dB. Le premier graphique affiche la variation du signal, le deuxième la variation de l'énergie des trois coefficients de haute fréquence de la transformée en ondelettes. Le seuil adaptatif de détection est présenté dans le deuxième graphique en pointillé et le seuil de fin sous forme de « tiret-point ». Nous pouvons observer que l'événement sonore est détecté avec précision même si le RSB est faible et que le silence en milieu de phrase est bien pris en compte.

3 Classification des évènements sonores

Les modèles de Markov cachés (HMM - Hidden Markov Models), les modèles de mélange de gaussiennes (GMM Gaussian Mixture Models)[3] et l'alignement temporel dynamique (DTW - Dynamic Time Warping) représentent les techniques de classification les plus utilisées dans le domaine de la recon-

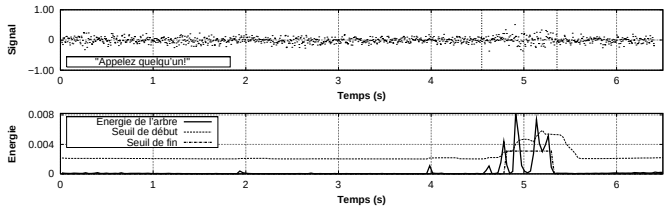

FIG. 3 – Exemple de détection de l'expression « Appelez quelqu'un » superposé au bruit HIS avec un RSB de 0 dB

naissance de la parole ou du locuteur.

Un système de classification statistique modélise chaque classe de signaux par une variable aléatoire qui est souvent une distribution de Gauss. La classification statistique est le calcul de la vraisemblance d'appartenance du signal à chacune des classes possibles, ce qui permettra de déterminer la classe d'appartenance la plus probable. Le calcul utilise une paramétrisation acoustique du signal et les modèles des classes à identifier.

Les modèles de mélange de distributions de Gauss (GMM) sont utilisés dans le cas de signaux complexes où il est nécessaire de considérer plus d'une variable aléatoire. La classification de sons à l'aide d'un modèle GMM comprend 2 étapes : une phase d'apprentissage du système sur un ensemble de fichiers supposés représentatifs d'une classe et, une deuxième phase, de vérification de l'appartenance d'un son quelconque à cette classe. La facilité d'emploi des GMM et leurs bonnes performances obtenues en reconnaissance des locuteurs représentent les principales raisons du choix de cette méthode pour le système proposé.

L'**apprentissage** a pour but d'estimer les paramètres des distributions de Gauss qui composent le modèle à partir des vecteurs acoustiques des sons composant la classe. L'apprentissage d'une classe se décompose en deux étapes successives : tout d'abord l'obtention de valeurs approximatives des paramètres des distributions de la classe par l'algorithme des K-moyennes, ensuite l'optimisation des valeurs de ces paramètres par un algorithme de type EM (**E**xpectation **M**aximisation).

La phase de **classification** permet de déterminer la classe la plus probable à partir du calcul de la vraisemblance pour chaque vecteur acoustique du signal. La vraisemblance d'un son constitué d'une suite temporelle de plusieurs vecteurs est la moyenne géométrique des vraisemblances de chacun de ses vecteurs. La classe de sons d'appartenance est celle pour laquelle la valeur de vraisemblance moyenne est maximale.

Paramètres acoustiques. L'apprentissage et la classification sont effectués non pas directement sur les signaux temporels mais sur des paramètres extraits de ceux-ci, parce que le signal temporel contient beaucoup d'informations redondantes. Le passage à une représentation fréquentielle du signal met en évidence les caractéristiques du signal.

Les paramètres acoustiques les plus fréquemment utilisés en reconnaissance de la parole sont les MFCC (Mel-Frequency Cepstral Coefficients), les LFCC (Linear Frequency Cepstral Coefficients) et les LPC (Linear Prediction Coefficients). Nous proposons l'utilisation de 3 paramètres acoustiques traditionnellement utilisés dans la segmentation parole/bruit/musique en addition aux paramètres MFCC. Ces 3 paramètres sont le nombre de passages par zéro (ZCR), le Roll-off Point (RF) et le barycentre spectral (Centroid).

Des coefficients cepstraux provenant de la transformée en ondelettes sont aussi étudiés. Ils sont obtenus grâce à quatre étapes de calcul : premièrement, calcul de la transformée en ondelettes, deuxièmement calcul de l'énergie des 6 derniers coefficients de la transformée, troisièmement application du logarithme décimal et enfin obtention du vecteur acoustique par calcul de la transformée en ondelettes inverse du vecteur d'énergies logarithmique. La dimension du vecteur acoustique est de 6 éléments et ces coefficients seront appelés DWTC.

Classification parole/sons et la classification des sons de la vie courante. La classification parole/sons a comme but d'identifier l'appartenance de l'événement sonore extrait lors de la phase de détection. Cette phase utilise 2 mélanges de distributions de Gauss pour modéliser la classe « parole » et la classe « son de la vie courante ».

La classification des sons est réalisée, elle aussi, avec un système à base de GMM. Chacune des 7 classes de sons de la vie courante est modélisée par un GMM.

4 Evaluation

Un corpus de sons de la vie courante a été réalisé à partir d'enregistrements effectués dans l'appartement de test et des sons de CDs commerciaux [4]. Dans le cas de la parole, un corpus d'expressions de détresse spécifiques à l'application médicale a été enregistré dans le studio de notre laboratoire.

Chaque étape du système a été validée individuellement à l'aide d'un corpus spécifique extrait du corpus de test.

4.1 Résultats de la détection

L'évaluation de l'algorithme proposé basé sur la transformée en ondelettes est présentée dans le Tableau 1. La première colonne indique le rapport signal sur bruit et les deux suivantes le taux d'égale erreur pour le bruit HIS et le bruit blanc. Pour chaque type de bruit quatre valeurs de RSB ont été envisagées : 0, +10, +20 et +40 dB. Le taux d'égale erreur (TEE) est le taux de détection manquées lu sur la courbe ROC (Receiver Operating Characteristic Curve) lorsque le taux de fausses alarmes (TFA) est égal au taux de détections manquées (TDM).

Pour analyser les résultats nous devons surtout analyser les performances pour le bruit HIS avec des valeurs faibles de RSB (l'environnement sonore réel). L'algorithme proposé, basé sur la transformée en ondelettes, procure des bonnes performances pour le bruit HIS : le TEE est de 0% lorsque le RSB est supérieur ou égal à 10 dB et le TEE est de 3.7% lorsque le RSB est égal à 0 dB. Ses performances dans le bruit blanc sont acceptables. Avec l'algorithme proposé, nous obtenons un TEE de 0% sur les 60 fichiers de la base de validation réelle (mélange réel des sons avec le bruit HIS) ce qui vient confirmer ces résultats.

Précisons que cet algorithme a non seulement de très bonnes performances en terme de TEE mais aussi que la précision temporelle de détection du début et de la fin du signal est très bonne. La précision du début de signal est en moyenne de 20 ms et est indépendante du RSB ; celle de fin de signal est d'environ 100 ms. Il faut tenir compte de ce que la taille de la trame

TAB. 1 – Taux d'égal erreur de l'algorithme de détection proposé

RSB	TEE	
	HIS [%]	Bruit blanc [%]
0 dB	3.7	6
10 dB	0	4
20 dB	0	0
40 dB	0	0

d'analyse est de 32 ms avec un pas de 16 ms.

4.2 Classification parole/sons

Le nombre de distributions de Gauss a été fixé à 24, valeur optimale obtenue à l'aide du critère BIC [5]. L'apprentissage des deux classes (parole et sons) est réalisé sur les sons purs et le test sur des sons mélangés avec le bruit HIS pour des valeurs de RSB de 0, 10, 20 et 40 dB. La validation est effectuée en utilisant un protocole d'évaluation croisée : l'apprentissage utilise à tour de rôle 80 % du corpus et le reste, les 20 % restant, sert pour le test.

Les performances de la classification sont évaluées par le calcul du taux d'erreur de classification (TEC) qui représente le rapport entre le nombre d'erreurs de classification et le nombre total de sons à identifier. Les résultats de classification parole/sons pour 16 paramètres MFCC et l'énergie normalisée sont présentés dans le tableau 2. Nous pouvons observer que le taux d'erreur reste inférieur à 4.5 % pour des RSB variant de 10 à 40 dB et atteint 22 % pour RSB=0 dB.

TAB. 2 – Taux d'erreur de classification parole/sons

RSB	TEC [%]		
[dB]	Global	Parole	Sons
0	22	29.9	8.6
10	4.2	5.9	3.1
20	4.4	9.1	1.6
40	4.5	9.6	1.4

4.3 Classification des sons de la vie courante

La classification des sons de la vie courante a été validée sur un corpus comportant 7 classes de sons. L'apprentissage a été effectué sur les sons purs et le test sur les sons mélangés avec le bruit HIS à des RSB variant entre 0 et 40 dB. Le nombre de distributions de Gauss a été fixé à 4, valeur déterminée avec le critère BIC comme précédemment.

Les performances de classifications sont constantes pour RSB$\geq$20 dB ; le TEC diminue pour des RSB au-delà de cette valeur : pour 16 paramètres MFCC couplés avec le nombre de passages par zéro, le Roll-off Point, le barycentre spectral et pour les 16 paramètres LFCC couplés avec l'énergie, le TEC est de 26.82 % pour RSB=+10 dB (Figure 4). Ces valeurs ne sont pas acceptables parce que le RSB dans l'environnement

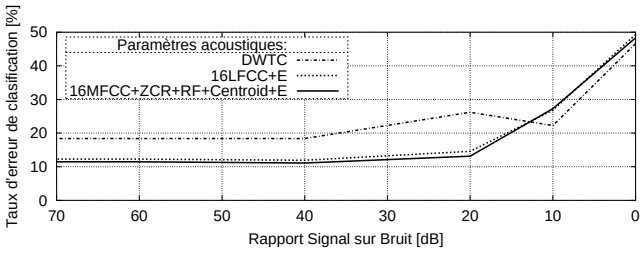

FIG. 4 – Erreur de classification dans le bruit HIS (apprentissage sur les sons purs)

de l'application varie entre 10 et 20 dB.

Les paramètres proposés DWTC ont des performances meilleures que les MFCC pour des RSB $\leq$ 10 dB et sont en nombre de 6 au lieu de 17 pour les autres.

5 Conclusions

Le cadre applicatif du système étudié est la télésurveillance médicale des personnes âgées ou des patients en convalescence. Le système utilise un réseau de 5 à 8 microphones installés dans un appartement d'étude à raison d'un au moins par pièce. Le but du système est de détecter une situation de détresse à partir de la surveillance sonore.

Pour le système proposé d'extraction de l'information du son, l'erreur de détection reste nulle lorsque le rapport signal sur bruit est plus élevé que +10 dB, et, par ailleurs la méthode de détection utilisée permet de garantir un positionnement précis du signal, typiquement 20ms de retard, et ne perturbe pas les étapes de classification. Dans les mêmes conditions de bruit, la classification parole/sons introduit un taux d'erreur inférieur à 5%. Nous pouvons donc conclure que le système de détection/classification peut être utilisé dans des conditions réalistes de fonctionnement avec un bruit modéré. La classification des sons de la vie courante est plus difficile à cause de la grande similarité entre les classes de signaux. Le taux d'erreur est de 12% pour un rapport signal sur bruit de +20 dB et de 26% à +10 dB, d'autres techniques devront être envisagées pour le réduire. Par ailleurs, une fusion de données avec d'autres capteurs est envisagée.

Références

[1] A. Dufaux, "Detection and recognition of impulsive sounds signals," Ph.D., Faculté des sciences de l'Université de Neuchatel, 2001.

[2] S. Mallat, *Une exploration des signaux en ondelette*, ser. ISBN 2-7302-0733-3. Palaiseau, France : Les Editions de l'Ecole Polytechnique, 2000.

[3] D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1, pp. 91–108, Jan. 1995.

[4] M. Vacher, D. Istrate, et J. F. Serignat, "Sound detection and classification through transient models using wavelet coefficient trees," en *EUSIPCO*, Vienne, Autriche, Sept. 2004, pp. 1171–1174.

[5] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, pp. 461–464, 1978.