Système de télésurveillance sonore pour la détection de situations de détresse

Dan Istrate, Michel Vacher, Jean-François Serignat, Laurent Besacier, Eric Castelli

Système de télésurveillance sonore pour la détection de situations de détresse

Remote Monitoring Sound System for Distress Situations Detection

Dan Istrate[ab,*], M. Vacher[a], J. F. Serignat[a], L. Besacier[a], E. Castelli[c]

[a] Laboratoire CLIPS – IMAG, UMR CNRS/INPG/UJF 5524, BP 53, 38041 Grenoble Cedex 9
[b]Laboratoire RMSE –ESIGETEL, 1, Rue du Port de Valvins,77215 Avon- Fontainebleau
[c] International Research Center MICA, 1, Dai Co Viet – Hai Ba Trung, Hanoi, Vietnam

Résumé

La télémédecine, et plus particulièrement la télésurveillance médicale, constitue aujourd'hui une solution pour pallier le manque de professionnels de santé face au fort accroissement de la population âgée en Europe. De plus, elle apporte à la fois une réduction des coûts d'hospitalisation et un meilleur confort au malade. Le système proposé, destiné à la surveillance de personnes âgées ou de malades chroniques à domicile, réalise la télésurveillance médicale à domicile à l'aide de capteurs sonores et médicaux en vue de la détection d'une situation de détresse. L'appartement d'expérimentation est équipé de capteurs médicaux (tensiomètre, oxymètre, balance, etc.), de capteurs de position à infrarouge et de microphones. L'originalité de ce système consiste à remplacer la surveillance vidéo, qui peut être mal perçue par les patients, par une surveillance sonore. Une division en deux étapes du système d'analyse sonore a été réalisée : la première étape a pour but la détection et l'extraction des événements sonores, elle est suivie de l'étape chargée de la classification sonore. Pour la première étape nous proposons un algorithme, basé sur la transformée en ondelettes, qui procure de bonnes performances en présence du bruit ambiant. L'étape suivante, la classification, utilise des modèles de mélange de distributions de Gauss pour classer l'événement sonore extrait parmi 7 classes de sons prédéfinies. L'algorithme de détection permet d'obtenir un taux d'égale erreur de 0% pour des rapports signal sur bruit supérieurs ou égaux à 10 dB et de 4% pour un rapport signal sur bruit inférieur à 10 dB. Le système d'analyse sonore proposé apportera des informations complémentaires au système classique de télésurveillance médicale auquel il sera couplé et contribuera à la fiabilité du système global.

Abstract

The telemedicine and medical remote monitoring in particular, today represents an effective solution to the health professional shortcomings facing to the increasing older population. In addition to the comfort of being at home, this system decreases the cost of long hospitalization. The proposed system achieves the home medical telesurveillance by means of microphones and medical sensors to detect a distress situation. This system is designed for elderly people at home or for the patient with chronic illness. The experimental apartment is equipped with medical sensors (tensiometer, oxymeter, balance, etc.), infrared position sensors, and acoustic sensors (1 microphone/room). The originality of the system comes from the replacement of the video surveillance with a sound surveillance. The sound analysis system is divided in two stages: firstly, the detection/extraction of the sounds is operated and secondly, a classification of these sounds in known classes takes place. A wavelet-based algorithm with good performance when applied in noisy environments is proposed. The acoustical classification step uses a Gaussian Mixture Models to classify the sounds according to the 7 predefined classes. The detection algorithm allows an equal error rate of 0 % for the signal to noise ratio superior or equal to 10 dB and 4 % for the 0 dB. The proposed system coupled with a classical medical telesurveillance system will bring extra information needed for the reliability of the global system.

Mots-clés : Analyse sonore ; Détection ; Modèle de mélange de distributions de Gauss ; Paramètre acoustique ; Transformée en ondelettes ; Télémédecine

Keywords: Sound Analysis, Detection, Gaussian Mixture Models; Acoustic Parameter; Wavelet Transform; Telemedicine

* Auteur correspondant
Adresse de correspondance : Résidence les Jarsines, 19Bis, Avenue de la Gare, 77250 Veneux les Sablons, Tel. 0681574378, E-Mail : dan.istrate@esigetel.fr

1. Introduction

Le nombre de personnes âgées étant en forte augmentation en Europe alors que le nombre des professionnels de la santé reste limité, la télésurveillance médicale à domicile devient une alternative efficace aux maisons de retraite médicalisées. Une solution possible est la *télémédecine*, l'utilisation des nouvelles techniques de l'information et de la communication pour des applications médicales. La télémédecine inclut les applications de télédiagnostic, télésurveillance [1], télé-opération, télé-éducation. Toutes ces tâches impliquent le partage de l'information, des données, de l'expertise et des services entre les professionnels de la santé.

Les technologies de l'information ont un rôle important dans l'évolution des services médicaux. Beaucoup d'applications ont déjà montré qu'en ce qui concerne le suivi médical des personnes âgées [2], [3], l'utilisation rationnelle de la télémédecine est une solution efficace aussi bien pour la qualité de service que pour les coûts d'hospitalisation.

Les problématiques de la télémédecine comprennent la surveillance des capteurs médicaux, la transmission des données, la compression des données sans pertes, la fiabilité des systèmes, etc. L'utilisation d'agents informatiques spécifiques [4] est une solution pour simplifier la complexité des logiciels et accroître leur fiabilité. La compression des données médicales est difficile parce qu'elle ne tolère pas des pertes de l'information [5]. Le choix du canal de transmission des données médicales est crucial, les solutions étudiées font surtout appel aux réseaux Ethernet, à la transmission de données à l'aide d'un terminal GSM (WAP) [6] et au câble de télévision [7].

La majorité des systèmes existants utilisent uniquement des capteurs médicaux (tensiomètre, oxymètre) et des capteurs de localisation (contacts de porte, infrarouge) pour la surveillance médicale [8], [9], [10]. D'autres systèmes font appel à la vidéo et au son, mais seulement en vue d'établir une communication entre le patient et le personnel soignant [11].

Cet article présente un système de détection et classification des sons de la vie courante permettant de développer une application de télésurveillance médicale sonore. Le système de télésurveillance doit couvrir toutes les pièces de l'appartement, y compris les toilettes, la salle de bains et la chambre. Si une caméra vidéo est installée dans chaque pièce, le patient peut avoir la sensation inconfortable d'être espionné. Non seulement un capteur sonore est moins indiscret mais il n'y a pas d'enregistrement continu du son : seul l'analyse en temps réel des 10 dernières secondes écoulées est effectuée par le système.

L'originalité du système proposé consiste dans l'utilisation du son comme source d'information, parallèlement aux autres capteurs. Le système d'analyse sonore détecte et identifie des sons de la vie courante comme les claquements de porte, les bris de verre, les sons de vaisselle, etc. Le but recherché consiste à détecter des situations de détresse telles qu'une chute ou un malaise partout dans l'appartement. Le remplacement des caméras vidéo par un système d'analyse du son simultanément sur plusieurs canaux permettra une analyse de l'environnement sonore de l'appartement en temps réel en vue de détecter une situation de détresse.

L'intimité de la personne est préservée parce que le son n'est jamais enregistré en continu; seul le dernier événement sonore détecté est enregistré et sera envoyé au centre de télésurveillance s'il est considéré comme un signe plausible d'alarme. Ce signal sera alors validé par le superviseur humain, qui pourrait décider de lancer une intervention d'urgence.

Pour des raisons de temps de calcul, sous la contrainte du temps réel, le système a été divisé en deux étapes : détection et classification. La deuxième section présente le système global de télésurveillance tandis que la troisième section décrit la base des sons utilisée pour la validation du système proposé. La quatrième section est consacrée à la détection des événements sonores et la cinquième section à la classification des sons de la vie courante. L'évaluation du système de télésurveillance sonore est traitée dans la sixième section. La dernière section présente la conclusion de cette étude.

2. Le système de télésurveillance

L'appartement de test (appelé Habitat Intelligent Santé) a une superficie de 30 m^2 et il est situé dans les locaux du laboratoire TIMC[1]. L'appartement est équipé de différents capteurs : capteurs volumétriques infrarouge, tensiomètre, oxymètre, balance, microphones. Le PC d'analyse sonore est relié à des microphones placés dans chacune des pièces : cuisine, couloir, séjour, douche et toilettes. Le système de télésurveillance complet est composé de deux ordinateurs qui échangent leurs informations par une connexion Ethernet (Figure 1).

Le PC d'analyse globale réalise la fusion de données entre les capteurs fixes et mobiles et l'ordinateur d'analyse du son. Le système d'analyse des capteurs de position et médicaux et une première réalisation de la

[1] Techniques de l'Imagerie, de la Modélisation et de la Cognition, UMR 5525

fusion de données avec le capteur sonore intelligent sont présentés en [12] et donc ils ne le seront pas présentés dans le cadre de cet article. L'ordinateur de fusion de données enverra une alarme ou non, en fonction des informations provenant à la fois du système d'analyse sonore et des autres capteurs.

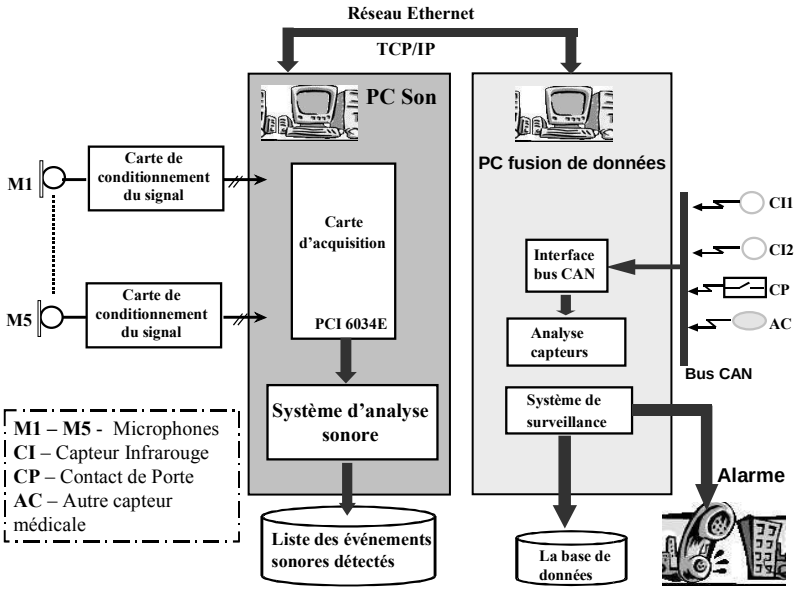

Figure 1 – Schéma du système de télésurveillance

2.1. Système d'analyse sonore

Le système d'analyse sonore a été décomposé en quatre modules pour des raisons de traitement temps réel, comme montré dans la Figure 2.

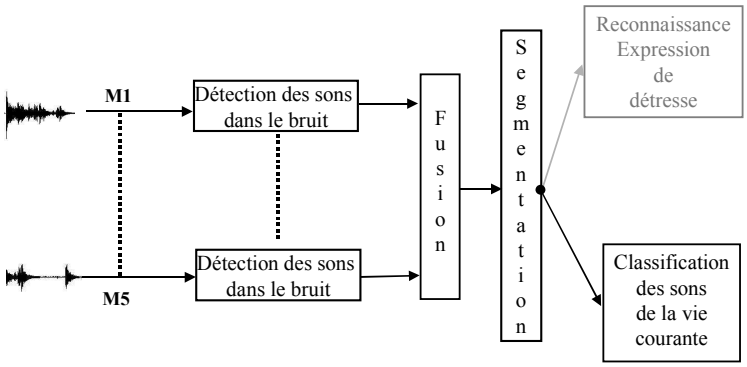

Figure 2 - Système d'analyse sonore

La *Détection* est lancée sur chaque canal sonore pour trouver et extraire les événements sonores. La localisation de la source audio est le résultat de la comparaison entre les rapports signal sur bruit de chaque canal. Le module de *Fusion* sélectionne le canal avec le meilleur rapport signal sur bruit dans le cas d'une détection simultanée sur plusieurs canaux.

Le module de *Segmentation*, basé sur des mélanges de distributions de Gauss (GMM), classe le signal extrait dans la classe *parole* ou dans la classe *sons de la vie courante* [13]. S'il s'agit de la parole, un système autonome de reconnaissance automatique de la parole est utilisé (moteur HMM Raphaël) [14]. Le modèle de langage du système Raphaël est un modèle statistique large vocabulaire (environ 20000 mots) élaboré à partir d'informations textuelles extraites du web par une technique mise au point par D. Vaufreydaz [15]. Il est optimisé pour l'application (appels de détresse) grâce à une méthodologie d'adaptation automatique, analogue à celle décrite dans [16]. Dans le cas contraire, le signal extrait est aiguillé vers la phase de classification des sons. Les modules de *Segmentation* et de reconnaissance de la parole, qui sont issus de l'état de l'art, ne seront pas étudiés au cours de cet article.

Le module de *Classification des sons* détermine la classe d'appartenance la plus probable pour l'événement sonore extrait parmi des classes prédéfinies.

Le système d'analyse sonore a été conçu pour répondre à plusieurs contraintes : l'analyse en temps réel des cinq canaux sonores, la large dynamique du signal, le fonctionnement permanent (24h/24h), la présence éventuelle de bruit non stationnaire de forte amplitude, la large variété des sons à identifier.

3. La base de test

Pour évaluer et valider le système de télésurveillance sonore, nous avons réalisé un corpus de sons. Ce corpus contient des enregistrements effectués dans le studio du laboratoire CLIPS (15% du CD [17]), des sons de la base *Sound Scene Database in Real Acoustical Environments* [18] (70% du CD [17]) et des enregistrements d'un CD commercial d'effets pour les films [19] (15% du CD [17]). Globalement, la base comprend 3354 fichiers sonores échantillonnés à 16 kHz et 44.1 kHz, elle représente 3 heures de signal.

3.1. La base de test de détection

La validation de l'algorithme de détection a nécessité la génération d'une base de test en mélangeant les sons utiles avec le bruit environnemental. Deux enregistrements ont été prévus pour chaque son : le premier est constitué par le mélange entre le son à détecter et le bruit, le second contient seulement le bruit. Chaque son et bruit ont été répétés trois fois. Chaque fichier a une durée de 25s, imposée par la taille des sons et par le temps d'initialisation de l'algorithme (5s). La base contient ≈ 2000 fichiers, 2 types de bruits : bruit blanc et bruit HIS (bruit de l'environnement de l'appartement de test) et 7 types de sons. Cette base a été réalisée pour quatre valeurs du RSB : 0, 10, 20 et 40 dB.

Pour valider le système en conditions réelles, nous avons enregistré un corpus de test spécifique dans l'appartement HIS (60 fichiers). Le RSB de ces fichiers varie entre 2 et 30 dB.

3.2. Les classes de sons

Les sons de la vie courante de notre application ont été divisés en 7 classes de sons. Les critères utilisés dans le choix des classes sont : la probabilité statistique d'apparition dans la vie courante, les sons critiques pouvant indiquer une situation de détresse (cris, chute), la durée des sons (sont considérés comme significatifs les sons courts et impulsionnels).

Les 7 classes de sons, présentées dans le Tableau I, peuvent donc être divisées en deux catégories :

- Sons **normaux** indiquant une activité usuelle de l'individu (claquement de porte, sonnerie de téléphone, sons de pas, sons de vaisselle) ;
- Sons **anormaux** qui peuvent signifier une situation de détresse pour le patient (bris de verre, cris).

Tableau I - Les classes de sons du système de télésurveillance

Classe de sons	Nbr. sons	Durée	Alarme
Claquement de porte (C1)	523	0.14–7.4 s	Non
Bris de verre (C2)	88	0.33-1.1 s	**OUI**
Sonnerie de téléphone (C3)	517	35ms-10s	Non
Sons de pas (C4)	13	1.4-5 s	Non
Cris (C5)	73	0.37-5.8 s	**OUI**
Vaisselle (C6)	163	0.13-1.4 s	Non
Serrure (C7)	200	24-117ms	Non

4. Détection

Pour la télésurveillance sonore, les performances de la détection sont très importantes car il est préférable d'éviter de perdre des événements sonores significatifs. Par contre, s'il y a beaucoup de fausses alarmes (détection en absence du signal) le module de *Classification des sons* risque d'être saturé.

La détection consiste à identifier le début et la fin des événements sonores dans un environnement bruité. Les deux hypothèses de la détection binaire sont :

$$\begin{cases} H_0 : & o(t) = & b(t) \\ H_1 : & o(t) = & s(t) + b(t) \end{cases} \quad \textbf{(1)}$$

où o(t) est le signal analysé, b(t) le bruit et s(t) le signal à détecter. Le principe de base de la détection est l'extraction d'un (ou plusieurs) paramètre(s) du signal d'entrée par une fonction suivie de la comparaison entre la séquence des valeurs obtenues et un seuil.

La détection du signal est un domaine très vaste qui inclut la détection des signaux numériques dans le bruit, la détection des signaux radar, la détection de l'activité vocale. Il y a plusieurs possibilités de choix pour la fonction de détection : l'énergie, la vraisemblance par rapport à un modèle statistique, les statistiques d'ordre supérieur du signal. La majorité des systèmes existants du domaine du traitement sonore détecte la parole (détection d'activité vocale VAD) et non pas des signaux impulsionnels [20] qui par ailleurs peuvent être très courts (jusqu'à 20 ms). La détection de la parole s'appuie sur des propriétés de la parole comme la fréquence fondamentale, les caractéristiques spectrales, les paramètres de prédiction linéaire (LPC). Le nombre de travaux liés à la détection des signaux impulsionnels est *réduit*. Dufaux (2001) a proposé trois algorithmes de détection des signaux impulsionnels avec de bonnes performances en présence de bruit blanc : un très simple, basé sur la variance de l'énergie du signal et deux autres algorithmes basés sur le filtrage médian conditionné de l'énergie du signal [21]. L'algorithme basé sur le filtrage médian conditionné sera utilisé dans notre étude comme l'algorithme de référence ; le seuillage est appliqué sur la différence entre l'énergie du signal et l'énergie filtrée avec le filtre médian conditionné [22].

Nos premières expériences nous ont montré que le bruit environnemental de l'appartement de test a des propriétés très différentes de celles du bruit blanc, ce qui conduit à une forte décroissance des performances de détection. Ce constat nous a amenés à proposer et optimiser un algorithme de détection bien adapté au bruit expérimental qui a pour caractéristique d'être basse fréquence et de contenir des sons impulsionnels provenant du voisinage de l'appartement expérimental.

4.1. Algorithme de détection basé sur la transformée en ondelettes

Comparée à la transformée de Fourier, la transformée en ondelettes est mieux adaptée aux signaux ayant des caractéristiques bien localisées dans l'espace temps-fréquence.

Tous les signaux x(t) peuvent être décomposés en une somme de fonctions $\Psi_{u,s}(t)$ pondérées par un facteur de poids et de localisation $\kappa_{u,s}$:

$$x(t) = \sum_{u,s} \kappa_{u,s} \Psi_{u,s}(t) \quad \textbf{(2)}$$

où u représente le décalage temporel et s le facteur d'échelle. Les fonctions $\Psi_{u,s}(t)$ sont à la base de la différence entre la transformée de Fourier à court terme (*analyse temps-fréquence*) et la transformée en ondelettes (*analyse temps-échelle*). En effet dans le cas de la transformée de Fourier, les sinusoïdes n'étant pas à support compact, les seuls paramètres sont la fréquence et la phase.

La transformée en ondelettes discrète (DWT) a une résolution fréquentielle et, respectivement, temporelle non-uniforme ; au contraire de la résolution fréquentielle, la résolution temporelle est meilleure dans les hautes fréquences et moins bonne dans les basses fréquences. Ces propriétés permettent une meilleure détection temporelle des signaux de haute fréquence, ce qui est le cas des signaux impulsionnels. La base de représentation de la transformée en ondelettes est générée par translation et dilatation de l'ondelette mère Ψ , qui dans notre cas est l'ondelette de Daubechies. Dans les applications de traitement du signal (débruitage, compression des signaux), les ondelettes de Daubechies sont choisies pour leurs propriétés (fonction lisse même pour un grand nombre de moments) ; nous avons choisi d'utiliser pour la détection des ondelettes de Daubechies avec 6 moments [23], [24].

La transformée en ondelettes discrète appliquée sur une fenêtre du signal délivre en sortie un vecteur de même taille. Ce vecteur a une structure pyramidale et il est composé de 10 coefficients de la transformée pour une fenêtre de calcul de 2048 échantillons. La structure des coefficients est présentée dans la Figure 3.

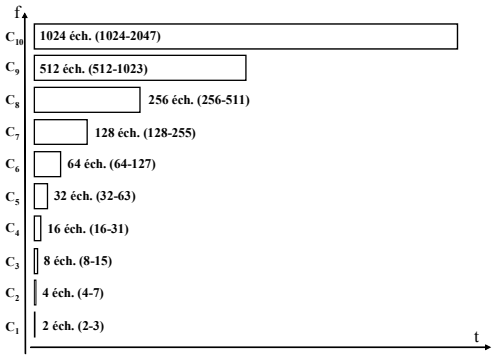

Figure 3 - Représentation des coefficients de la transformée en ondelettes pour 2048 échantillons

L'algorithme proposé (Figure 4) calcule l'énergie des trois coefficients de haute fréquence de la transformée en ondelettes (les coefficients du plus grand ordre qui sont composés de 1024, 512 et 256 composants), parce que les sons à détecter sont caractérisés par ces coefficients de haute fréquence. La taille de la fenêtre d'analyse est de 128 ms (2048 échantillons), elle est imposée par les contraintes de temps réel. Une meilleure résolution temporelle étant indispensable, le vecteur de sortie de la transformée sera découpé en 4 trames d'analyse de 32 ms chacune. La trame d'analyse a donc, une taille de 32 ms avec un recouvrement de 16 ms (50%).

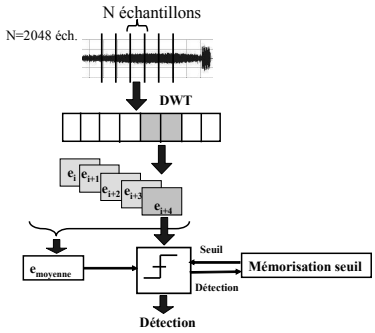

Figure 4 - La diagramme de l'algorithme de détection basé sur la transformée en ondelettes

La détection sera déterminée grâce à un seuil appliqué sur la somme des énergies des trois ondelettes les plus hautes de la trame. Le seuil est adaptatif (équation (3)) et il dépend de la moyenne de N valeurs de l'énergie μ_E (dans cette étude 40 valeurs sont utilisées pour la représentation statistique). Une phase d'apprentissage est effectuée pour le réglage de α qui dépend du niveau du signal acquis ; il est réglé à 20 % de la valeur maximale de l'énergie des coefficients de la transformée en ondelettes.

$$S = \alpha + 1.2\mu_E \quad \textbf{(3)}$$

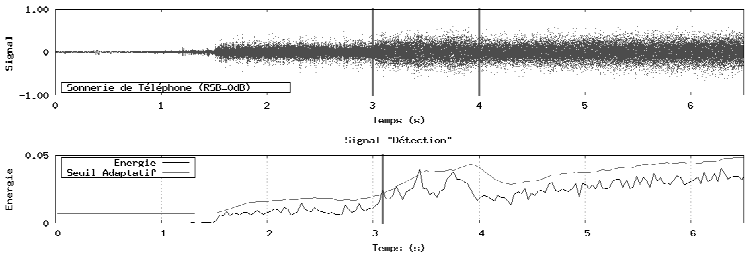

Figure 5 - Exemple de détection d'une sonnerie de téléphone en présence d'un bruit d'écoulement d'eau avec un RSB=0dB

Un exemple de détection est présenté dans la Figure 5 avec une sonnerie de téléphone, commençant à l'instant t=3s, en présence d'un bruit d'écoulement d'eau selon un rapport signal sur bruit (RSB) de 0dB (se reporter à la fenêtre du haut). Dans la fenêtre du bas de la même figure, a été présentée la somme des énergies des trois coefficients supérieurs de la transformée en ondelettes en trait plein, ainsi que le seuil adaptatif qui apparaît en pointillé. Nous observons que la sonnerie de téléphone est correctement détectée par cet algorithme.

La procédure de détection de la fin du signal débute au moment de la détection du début avec la mémorisation de la valeur du seuil adaptatif. On considère que la fin du signal est détectée à partir de l'instant où l'énergie se trouve en dessous du seuil pour un nombre fixé M de fenêtres d'analyse. Pour M, une valeur de 16 trames (0.25 s) a été choisie en vue de s'affranchir des silences existant entre les mots, et de permettre l'utilisation de l'algorithme dans le cas de la parole. La valeur de M a été obtenue après l'étude statistique du corpus de parole française BRAF100 qui contient l'enregistrement de 100 locuteurs, 10 000 phrases, 20 000 mots et une durée de 28 heures. Dans ce corpus nous avons en moyenne 3 mots/s, avec une durée moyenne d'un mot de 0,33 s donc un silence moyen entre les mots de 0,2 s.

4.2. Validation de l'algorithme

4.2.1 Evaluation des performances de la détection

Le taux de détections manquées (TDM) et le taux de fausses alarmes (TFA) sont évalués pour caractériser les performances de la détection sur la base de test. Les formules définissant ces deux taux sont :

$$TDM = \frac{\text{Nbr. détections manquées}}{\text{Nbr. de détections}} \quad \textbf{(4)}$$

$$TFA = \frac{\text{Nbr. fausses alarmes}}{\text{Nbr. fausses alarmes} + \text{Nbr. détections}} \quad \textbf{(5)}$$

Une détection est considérée comme *fausse* quand il y a une détection en absence d'événement. Une détection est considérée comme *manquée* quand le système ne détecte rien dans un intervalle commençant 0.5s avant l'événement et allant jusqu'à la fin du signal. Une détection qui se produit dans ce même intervalle est par contre une *bonne* détection.

Pour comparer les algorithmes de détection, le taux d'égale erreur (TEE) est calculé à partir des courbes ROC (Receiver Operating Characteristics, TDM=f(TFA)), il est défini comme étant la valeur du TDM pour lequel TDM=TFA (l'intersection entre la première bissectrice et la courbe ROC).

4.2.2 Les performances de l'algorithme de détection proposé

L'évaluation de l'algorithme de référence et de l'algorithme proposé basé sur la transformée en ondelettes est présentée dans le Tableau II. La première colonne indique l'algorithme étudié, la deuxième 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 0, 10, 20 et 40 dB ont été envisagées.

Pour analyser les résultats nous devons comparer les performances pour le bruit HIS avec des valeurs réduites du RSB (l'environnement sonore réel). L'algorithme de référence n'est pas convenable parce que le TEE dépasse 10% pour des valeurs du RSB inférieures ou égales à 20 dB. L'algorithme proposé, basé sur la transformée en ondelettes, procure les meilleures 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 décrite à la fin de la section décrivant la base de test, ce qui vient confirmer ces résultats.

Tableau II - Les performances de l'algorithme de l'état de l'art comparées avec celles de l'algorithme proposé

Algorithme de détection	RSB [dB]	TEE	
		HIS [%]	Bruit blanc [%]
Basé sur la DWT	0	**3.7**	6
	10	**0**	4
	20	**0**	0
	40	**0**	0
Référence	0	65	30
	10	36	0
	20	10	0
	40	0	0

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. Le Tableau III présente la

précision de détection du début et de la fin du signal pour les quatre valeurs de RSB. Nous observons que la précision du début est en moyenne de 20 ms, indépendante du RSB. Il faut tenir compte que la taille de la trame d'analyse est de 32 ms avec un pas de 16 ms. La détection avec précision du signal est importante pour la phase suivante de classification. Cette valeur a été choisie en tenant compte de la durée minimale des sons du corpus (24 ms). L'influence d'une mauvaise extraction de l'événement sonore est étudiée dans la section suivante.

Tableau III - La précision temporelle (ms) de détection du début et de la fin du signal

RSB Type	0 dB	10 dB	20 dB	40 dB
Début	22	18	19	23
Fin	159	114	100	81

La Figure 6 présente la courbe ROC de l'algorithme de détection basé sur la transformée en ondelettes en présence du bruit HIS pour les quatre valeurs de RSB ; pour des RSB ≥ 10 dB il s'agit des courbes qui coïncident avec les axes.

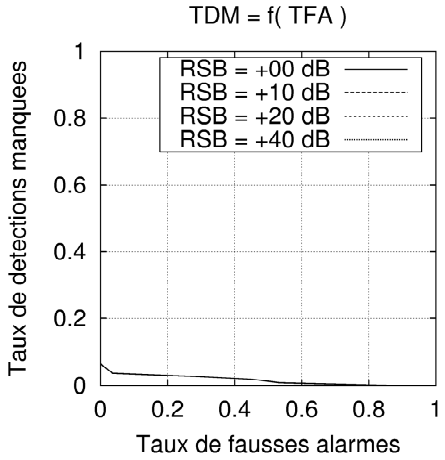

Figure 6 - La courbe ROC de l'algorithme de détection pour le bruit HIS

5. Classification des sons de la vie courante

La « *Classification »* désigne l'identification du son ou de sa source à partir de quelques aspects particuliers. Dans notre cas il s'agit d'identifier une classe de sons et pour cela nous pouvons étudier des techniques du domaine de la reconnaissance des formes telles que le modèle de mélange de distributions de Gauss (GMM) [25], le modèle de Markov caché (HMM) [26], l'alignement temporel dynamique (DTW) [27], les réseaux de neurones. Les études sur la reconnaissance des sons environnementaux sont peu nombreuses et dans une phase préliminaire. Woodard [28] utilise un HMM pour classifier seulement 3 classes de sons en l'absence de bruit. Un système de classification qui compare le spectre normalisé avec des spectres appris est proposé dans [29] pour 3 classes de sons. Une comparaison entre réseaux de neurones, DTW et quantification vectorielle est proposée dans [30].

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. Les HMM sont plus complexes, avec un temps de calcul plus important et moins adaptés pour la classification des signaux courts. La classification à base de GMM comprend 2 étapes : une phase d'*apprentissage* et une phase de *classification*.

La phase d'*apprentissage* est initiée pour chaque classe sonore du corpus et a comme but l'estimation des paramètres des distributions de Gauss de chaque modèle. Cette phase commence par l'algorithme de K-moyennes qui est suivi de 20 itérations de l'algorithme EM (Expectation-Maximization).

La phase de classification d'un son détecté se réduit au calcul de la vraisemblance du signal par rapport à chacune des classes de son. Finalement, la classe reconnue est celle ayant la plus grande valeur de vraisemblance.

5.1. 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).

Le calcul des paramètres MFCC se décompose en 3 étapes : transformée de Fourier Rapide, logarithme de l'énergie de 24 bandes de filtres Mel, transformée inverse en cosinus discret. Les filtres Mel sont des filtres triangulaires qui utilisent une échelle fréquentielle non linéaire qui tient compte des particularités de l'oreille humaine [31]. Ce sont des filtres uniformes en fréquence qui sont utilisés pour obtenir les paramètres LFCC.

Nous proposons l'utilisation de 3 paramètres acoustiques traditionnellement utilisés dans la segmentation parole/bruit/musique complétés avec les paramètres MFCC. Ces 3 paramètres sont le nombre de passages par zéro (ZCR), le Roll-off Point et le barycentre des énergies (Centroid). La première et la deuxième dérivée des paramètres acoustiques (appelées respectivement Δ et $\Delta\Delta$) sont calculées pour introduire la variation temporelle du signal dans la modélisation GMM.

Le *nombre de passages par zéro (ZCR)* représente le nombre de passages par zéro du signal temporel dans la fenêtre d'analyse et généralement, il indique la fréquence dominante du signal dans la fenêtre.

Le *Roll-off point (RF)* est la fréquence au-dessous de laquelle se situe 95 % de l'énergie du signal. On peut dire que c'est un indice de répartition du spectre de puissance du signal. Le RF est plus grand pour les signaux ayant un spectre haute fréquence. Le RF est la solution de l'équation (6) avec $\Theta = 0{,}95$.

$$\sum_{k<RF} X[k] = \Theta \sum_{k} X[k] \quad \textbf{(6)}$$

Le *Centroid* spectral est la valeur de la fréquence partageant le spectre en deux parties d'égale énergie : basse fréquence/haute fréquence. Le centroide est la solution de l'équation (6) en fixant la valeur du paramètre Θ à 0,5.

Les paramètres acoustiques les plus pertinents pour les classes de sons ont été identifiés à l'aide du critère de Fisher (FDR Fisher Discriminant Ratio) [32]. Dans l'équation (7) la moyenne du paramètre x pour la classe i est $\overline{x}[i]$, la variance du paramètre x pour la classe i est $Var(x)[i]$ et k est le nombre de classes.

$$FDR = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k}\left(\overline{x}[i] - \overline{x}[j]\right)^2}{\sum_{i=1}^{k} Var(x)[i]} \quad (7)$$

5.2. Résultats de classification

5.2.1 Nombre de distributions de Gauss du modèle

Parmi les classes de sons, il y a des classes qui peuvent contenir un nombre assez faible d'échantillons, l'utilisation d'un nombre trop important de distributions de Gauss ne conduira donc pas nécessairement à une meilleure modélisation. Parmi les critères de sélection de la meilleure modélisation d'une classe de sons, nous avons considéré le critère BIC (Bayesian Information Criterion) [33]. Ce critère sélectionne le nombre optimal de distributions de Gauss en maximisant la vraisemblance intégrée, conformément à l'équation (8).

$$BIC_{m,k} = -2.L_{m,k} + \nu_{m,k} \ln(n) \quad (8)$$

où $L_{m,k}$ est le logarithme du maximum de vraisemblance, m est le modèle et k le nombre de composants du modèle, $\nu_{m,k}$ est le nombre de paramètres libres du modèle m et n le nombre de trames. La valeur minimale du critère BIC indique le modèle optimal.

Le critère BIC a été calculé pour la classe de sons comportant le nombre le plus réduit de fichiers. Les résultats obtenus pour 16MFCC sont présentés dans le Tableau IV. En analysant ce tableau nous trouvons que la valeur optimale se trouve entre 3 et 5 distributions de Gauss ; un nombre de 4 distributions de Gauss a donc été choisi pour le reste des expérimentations.

Tableau IV – Critère BIC pour 2 à 8 distributions de Gauss

N°. Gauss	2	**3**	**4**	**5**	8
BIC	11043	**10752**	**10743**	**10757**	13373

5.2.2 Résultats de classification

Les résultats de l'évaluation de la phase de classification sont présentés dans le Tableau V en utilisant une fenêtre de calcul de paramètres acoustiques de 16 ms avec un recouvrement de 8 ms. Le protocole d'évaluation est de type « leave-one out » : chaque signal sonore (parmi les 1577) est extrait de la base d'apprentissage et il est utilisé pour le test.

La première colonne du Tableau V présente le type des paramètres acoustiques, la deuxième leurs nombre et la troisième le taux moyen d'erreur de classification (TEC). Nous pouvons observer que les meilleurs résultats sont obtenus avec les paramètres de type MFCC couplés avec des nouveaux paramètres comme le nombre de passage par zéro, le Roll-off point et le Centroid. En rajoutant à ces paramètres la première et la deuxième dérivée, nous obtenons avec 60 paramètres les meilleures performances.

Le critère de Fisher montre que les 3 paramètres MFCC couplés avec ZCR, RF et Centroid sont les plus pertinents. L'élimination des paramètres considérés comme non pertinents induit une réduction non importante des performances de classification ; une diminution du nombre des paramètres acoustiques de 20 à 6 implique une réduction des performances absolues de seulement 4.5 %.

Tableau V - Les performances de la classification des sons

Paramètres	Nbr	TEC [%]
Δ,ΔΔ(16MFCC+Energie+ZCR+RF +Centroide)	60	8.7
16MFCC+Energie+ZCR+RF+ Centroide	**20**	**11.4**
16LFCC+Energie	17	12.2
16LFCC+ZCR+RF+Centroide	19	12.7
16MFCC+Energie	17	15.2
3MFCC+ZCR+RF+Centroide	**6**	16.1

L'influence de la précision d'extraction des événements sonore a été évaluée par l'étude de trois cas :

1. Détection manuelle des événements (durée réelle)
2. Détection automatique du début de l'événement avec durée fixe de 7s (durée maximale des événements à détecter)
3. Détection automatique du début et de la durée du signal

Les résultats obtenus, Tableau VI, montrent une influence négative importante de la précision d'extraction du signal. Dans le cas de l'extraction avec durée fixe le signal envoyé à l'étape de classification contient des parties importantes de bruit environnemental sans signal utile qui fausse la classification qui est obtenue à la suite d'une moyenne de vraisemblances.

Tableau VI - Evaluation de l'influence de la précision de détection sur le système

	TEC [%] RSB$\in$(10,20)dB
Référence (détection manuelle) TDM=0% ; TFA=0 %	**21.5** %
Détect. aut. et durée fixe TDM=1% TFA = 1%	67.8 %
Détect. automatique TDM=1% TFA=1%	**27.7** %

Tableau VII présente la matrice de confusion pour les meilleurs paramètres acoustiques (16MFCC+Energie+ZCR+RF+Centroide). Nous observons que les classes de sons difficiles à modéliser sont : « Claquement de porte (C1) » et « Vaisselle (C6) ». Les classes qui attirent plus les sons des autres classes sont « Claquement de porte (C1) », « bris de verre (C2) » et « son de pas (C4) » ; ces classes présentent des caractéristiques communes avec toutes les autres.

Tableau VII - Matrice de confusion pour la classification des sons

	C1	C2	C3	C4	C5	C6	C7
C1	**458**	26	0	37	0	2	0
C2	3	**85**	0	0	0	0	0
C3	17	0	**499**	0	0	1	0
C4	2	0	0	**11**	0	0	0
C5	1	0	1	0	**71**	0	0
C6	59	17	0	0	1	**86**	0
C7	2	12	0	0	0	0	**186**

6. Evaluation du système global

L'évaluation globale du système implique la division des sons en deux classes : les sons pouvant générer une alarme et les sons normaux.

Les cas possibles après la phase de détection sont : **bonne détection** (BD), **fausse détection** (FD), **détection manquée** (DM). Les événements détectés (BD et FD) constituent l'entrée de l'étape de classification. Une partie des événements détectés sont sans conséquence (DM S) parce qu'ils représentent des sons normaux.

L'étape de classification (voir Figure 7) peut donner lieu à : des **bons événements** (BE), des **faux événements** (FE) et des **événements manqués** (EM). Comme pour la première étape une partie des événements manqués sont sans conséquence (EM S).

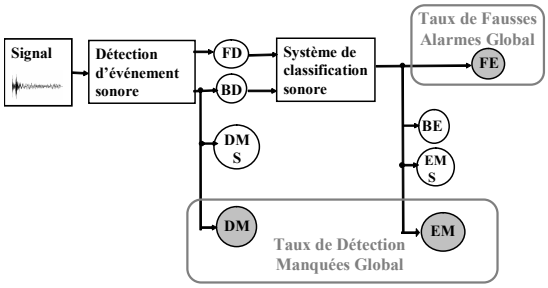

Figure 7 - Taux de détections manquées global et taux de fausses alarmes global

Le taux de détections manquées global dépend des détections manquées et des événements manqués. Le Taux de fausses alarmes global tient compte des faux événements.

Tableau VIII - Les performances globales du système de télésurveillance sonore

TDM Global	TFA Global
3 %	12 %

Le taux de détections manquées global est de 3 % pour le système entier (Tableau VIII). Ce taux est acceptable car le système global de fusion utilise aussi les données fournies par les autres capteurs.

Le taux de fausses alarmes global est de 12 %. Pendant une « journée normale » d'une personne âgée, le nombre moyen des événements sonores peut être estimé à une valeur d'environ 100 ; donc nous pouvons avoir 12 fausses alarmes par jour. Une partie de ces fausses alarmes peut être éliminée par la fusion avec les informations fournies par les capteurs de position infrarouge, d'activité (accéléromètre) et d'autres capteurs médicaux (tensiomètre, oxymètre, balance).

7. Conclusions et perspectives

Les principaux résultats de cette étude peuvent être résumés ainsi :

- un algorithme de détection sonore a été proposé et validé ; cet algorithme est basé sur la transformée en ondelettes et a de bonnes performances même en présence de bruit fort.
- un système de classification des sons de la vie courante a été présenté ; de nouveaux paramètres acoustiques ont été utilisés conjointement aux paramètres classiques
- une méthodologie d'évaluation des systèmes de télésurveillance médicale a été proposée.

L'algorithme de détection a été comparé avec un algorithme de référence, il permet la détection d'un événement sonore avec une erreur de 4 % pour un rapport signal sur bruit de 0 dB. La classification des sons de la vie courante est réalisée avec un système statistique de type GMM. Des paramètres acoustiques non

conventionnels ont été rajoutés aux paramètres classiques pour obtenir une amélioration des performances. Le système global présenté conduit à un taux de détections manquées de 3 % qui pourrait être diminué en fusionnant les données du système avec celles des capteurs de position infrarouge et celles des capteurs médicaux.

Le système a été validé module par module et en ensemble en utilisant des corpus sonores obtenus par le mélange simulé des signaux utiles réels avec le bruit de l'appartement de test. Une validation des résultats a été effectuée par l'utilisation d'un corpus de taille réduite enregistré dans l'appartement de test (mélange réel). La validation en conditions réelles ne pourra être envisagée qu'avec l'équipement d'appartements dans des maisons de retraites.

Ce processus d'extraction d'information du son peut être utilisé pour d'autres applications comme la classification des documents multimédia ou la surveillance de la sécurité des locaux.

Remerciements

Ces recherches ont été financées par le Ministère Français de la Recherche, dans le cadre du projet ACI Santé, DESDHIS (Détection de Situations de Détresses en Habitat Intelligent Santé). Ce projet entre dans le cadre d'une collaboration entre les laboratoires CLIPS (UMR CNRS-INPG-UJF 5524) et TIMC (UMR CNRS-INPG-UJF 5525).

Références Bibliographiques :

[1] I. Korhonen, J. Parkka, et M. V. Gils, « Health monitoring in the home of the future », *IEEE Eng. Med. Biol.Mag*., pp. 66–73, mai 2003.

[2] R. L. Bashshur, « State-of-the-art telemedicine & telehealth: Ch.1 - telemedicine and health care », *Telemedicine Journal and e-Health*, vol. 8, no. 1, pp. 5–12, 2002.

[3] P. A. Jennett, L. A. Hall, D. Hailey, A. Ohinmaa, C. Anderson, R. Thomas, B. Young, D. Lorenzetti et R. E. Scott, « The socio-economic impact of telehealth. A systematic review », Journal of Telemedicine and Telecare, vol. 9, no. 6, pp. 311–320, 2003.

[4] V. D. Mea, « Agents acting and moving in healthcare scenario - a paradigm for telemedical collaboration », *IEEE Trans. Inform. Technol. Biomed.*, vol. 5, no. 1, pp. 10–13, mars 2001.

[5] Z. Lu, D. Y. Kim et W. A. Pearlman, « Wavelet compression of ECG signals by set partitioning in hierarchical trees algorithm », *IEEE Trans. Biomed. Eng.*, vol. 47, pp. 849-856, 2000.

[6] K. Hung et Y. T. Zhang, « Implementation of a WAP-Based telemedicine system for patient monitoring », *IEEE Trans. Inform. Technol. Biomed.*, vol. 7, no. 2, pp. 101–107, juin 2003.

[7] R. G. Lee, H. S. Chen, C. C. Lin, K. C. Chang et J. H. Chen, « Home telecare system using cable television plants - an experimental field trial », *IEEE Trans. Inform. Technol. Biomed*., vol. 4, no. 1, pp. 37–44, mars 2000.

[8] M. Takizawa, S. Sone, K. Hanamura et K. Asakura, « Telemedicine system using computed tomography van of high-speed telecommunication vehicle », *IEEE Trans. Inform. Technol. Biomed.*, vol. 5, no. 1, pp. 2–9, mars 2001.

[9] J. Reina-Tosina, L. Roa et M. Rovayo, "NEWBET: telemedicine platform for burn patients », *IEEE Trans.Inform. Technol. Biomed.*, vol. 4, no. 2, pp. 173–177, juin 2000.

[10] E. Jovanov, A. D. Lords, D. Raskovic, P. G. Cox, R. Adhami et F. Andrasik, « Stress monitoring using a distributed wireless intelligent sensor system », *IEEE Eng. Med. Biol. Mag.*, pp. 49–55, mai 2003.

[11] P. Varady, Z. Benyo et B. Benyo, « An open architecture patient monitoring system using standard technologies », *IEEE Trans. Inform. Technol. Biomed.*, vol. 6, no. 1, pp. 95–98, mars 2002.

[12] G. Virone, D. Istrate, M. Vacher, J. F. Serignat, N. Noury et J. Demongeot, «First Steps in Data Fusion between a Multichannel Audio Acquisition and an Information System for Home Healthcare», *IEEE Engineering In Medicine And Biology Society Conference*, Cancun, Mexique, 13-15 septembre 2003, pp.1364-1367.

[13] D. Istrate, M. Vacher et J. F. Serignat, «Détection et classification des sons : application aux sons de la vie courante et à la parole», GRETSI 2005, Louvain-la-Neuve, Belgique, 6-9 septembre 2005, pp. 485-488.

[14] M. Akbar et J. Caelen, « Parole et traduction automatique : le module de reconnaissance RAPHAEL », COLING-ACL'98, Montréal, Quebec, vol.2, p. 36-40.

[15] D. Vaufreydaz, J. Rouillard, M. Akbar, « Internet Documents : a Rich Source for Spoken Language Modelling », IEEE Workshop ASRU'99, Keystone-Colorado, USA, décembre 1999, pp. 277-281.

[16] V.-B. Le et L. Besacier, « First steps in fast acoustic modeling for a new target language. Application to Vietnamese », IEEE ICASSP 2005, Philadelphia, USA, 19-23 mars, 2005, Vol. 1, pp.821-824.

[17] D. Istrate, « Base de données. Sons de la vie courante », CLIPS-IMAG Equipe GEOD, www-clips.imag.fr, Grenoble, France, novembre 2001.
[18] R. W. C. Partnership, « CD - Sound scene database in real acoustical environments », http://tosa.mri.co.jp/soundb/indexe.htm, Tokio, Japan, 1998 2001.
[19] S. Sciascia, « CD - bruitages - vol.3 », Paris, France, 1992.
[20] M. Marzinzik et B. Kollmeier, « Speech pause detection for noise spectrum estimation by tracking power envelope dynamics », *IEEE Trans. Speech Audio Processing*, vol. 10, no. 2, pp. 109–118, fevrier 2002.
[21] A. Dufaux, « Detection and recognition of impulsive sounds signals », Ph.D. dissertation, Faculté des Sciences de l'Université de Neuchatel, 2001.
[22] A. Dufaux, L. Besacier, M. Ansorge, A. Pellandini, « Automatic sound detection and recognition for noisy environment », EUSIPCO 2000, Tampere, Finland, 4-8 septembre, 2000.
[23] S. Mallat, *Une exploration des signaux en ondelettes*, ISBN 2-7302-0733-3. Palaiseau, France: Les Editions de l'Ecole Polytechnique, 2000.
[24] P. L. Dragotti et M. Vetterli, « Wavelet footprints: Theory, algorithms, and applications », *IEEE Trans. Signal Processing*, vol. 51, no. 5, pp. 1306–1323, mai 2003.
[25] D. A. Reynolds and R. C. Rose, « Robust text-independent speaker identification using gaussian mixture speaker models », *IEEE Trans. Speech Audio Processing*, vol. SAP-3, no. 1, pp. 72–83, janvier 1995.
[26] L. R. Rabiner et B. H. Juang, *Fundamentals of speech recognition*, ISBN 0-13-015157-2. Prentice Hall PTR, New Jersey, USA, 1993.
[27] C. S. Myers et L. R. Rabiner, « A comparative study of several dynamic time-warping algorithms for connected word recognition », *The Bell System Technical Journal*, 60(7), pp.1389–1409, 1981.
[28] J. P.Woodard, « Modeling and classification of natural sounds by product code hidden markov models », *IEEE Trans. Signal Processing*, vol. 40, no. 7, pp. 1833–1835, juillet 1992.
[29] G. Papadopoulos, K. Efstathiou, Y. Li et A. Delis, « Implementation of an intelligent instrument for passive recognition and two-dimensional location estimation of acoustic targets », *IEEE Trans. Instrum. Meas.*, vol. 41, no. 6, pp. 885–890, juin 1992.
[30] M. Cowling et R. Sitte, « Analysis of speech recognition techniques for use in a non-speech sound recognition system » in *Digital Signal Processing for Communication Systems*, Sydney-Manly, Australia, janvier 2002.
[31] S. B. Davis et Mermelstein, « Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences », IEEE Transactions on Audio, Speech and Signal Processing, 28(4), pp.357-366, 1980.
[32] D. Kil et F. Shin, *Pattern Recognition and Prediction with Applications to Signal Characterization*. AIP Press Woodbury, New York, 1996.
[33] G. Schwarz, « Estimating the dimension of a model », Annals of Statistics, vol. 6, pp. 461–464, 1978.