# Capteur sonore couplé avec capteur thermique pour la détection du nombre de personnes et des situations de détresse

Sami Boutamine[1], Dan Istrate[1], Jérôme Boudy[2], Adrian NICOLICEA[3]
[1]Sorbonne Universités, Université de Technologies de Compiègne, BMBI UMR7338, France
[2]Télécom Sud Paris, SAMOVAR-ARMEDIA UMR 5157 Evry, France
[3]Université de Craiova, Craiova, Roumanie
sami.boutamine@utc.fr

***Mots-clés: reconnaissance des sons, reconnaissance du locuteur, segmentation en locuteurs, GMM, i-Vector, LFFC, MFFC.***

## I. INTRODUCTION

Aujourd'hui les technologies au service des espaces intelligents ne cessent de se développer et sont en interaction avec les objets de la vie courante. Ils exploitent des signaux vidéo, des signaux audio et des données environnementales afin de permettre la localisation des personnes, la reconnaissance des gestes, etc.

Le projet FUICoCAPs entre dans cette catégorie de technologies, il se base sur le développement des capteurs à faible coût fournissant des informations enrichies sur le comportement des personnes dont le but est d'automatiser et de réduire la consommation de l'éclairage, de la ventilation, du chauffage, etc.

Le système proposé par le projet CoCAPs (*Figure1*) se décompose en trois modules :

1. Récupération et l'analyse des données des capteurs (son, image, environnement),
2. Fusion des résultats obtenus,
3. Système de décision

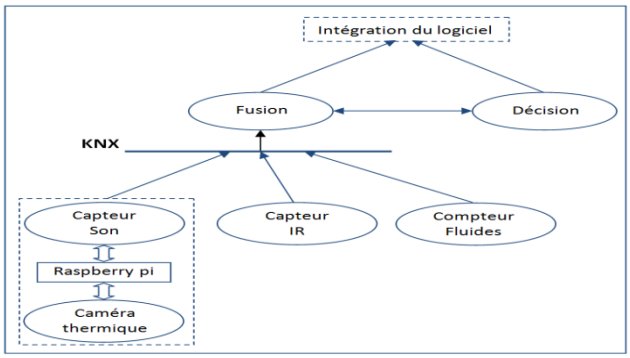


Figure 1. Architecture du système du projet CoCAPs.

## II. SYSTEME D'ANALYSE SONORE

L'objectif est multiple, à savoir à la fois la détection du nombre de personnes, et la détection des situations de détresse et d'activités en se basant sur un capteur sonore couplé avec capteur thermique multipoint.

### 2.1. Analyse de l'environnement sonore

Le système d'analyse sonore proposé est présenté dans la figure 2 et a pour objectifs la segmentation du signal sonore recueilli en segments correspondant aux différents locuteurs présents (reconnaissance du locuteur principal), la reconnaissance des sons de la vie courante et la reconnaissance d'ordres vocaux.

La segmentation en locuteurs et la reconnaissance du locuteur pourront être très utiles pour la détection du nombre de personnes et de leurs activités.

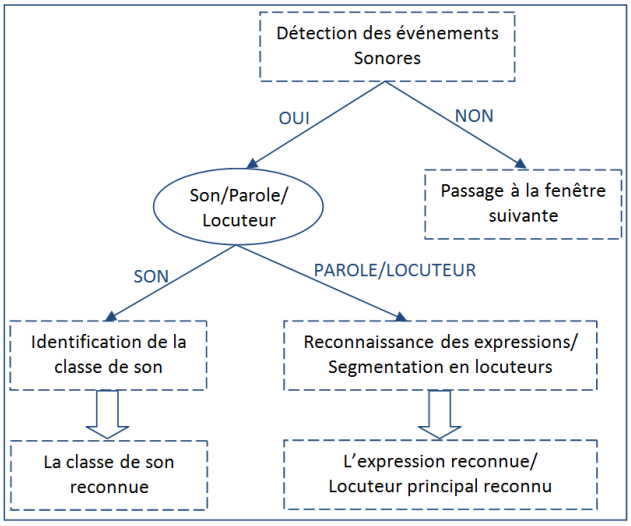


Figure2 : Système d'analyse sonore.

### 2.2. Reconnaissance du locuteur

Le fonctionnement du système de reconnaissance du locuteur [4] décrit en figure 3, exploite des fichiers audio, et s'appuie sur trois étapes principales : le calcul des vecteurs acoustiques, la création des modèles des locuteurs ou des sons et la reconnaissance du locuteur principal.

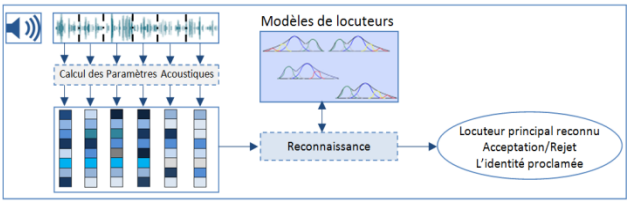


Figure 3 : Système de reconnaissance du locuteur.

Dans ce domaine on peut citer 3 tâches principales :

1. L'identification du locuteur : la détection de l'identité d'une personne parmi N personnes connues à l'avance par le système.
2. La vérification du locuteur : la vérification si un échantillon de voix correspond à une identité proclamée connue à l'avance par le système.
3. Le suivi du locuteur : la détection à partir d'un enregistrement audio où il y a plusieurs personnes qui parlent, les segments appartenant à certains locuteurs connus à l'avance par le système [7].

On remarque que la reconnaissance automatique du locuteur nous permet de détecter le nombre de personnes tout en basant sur les modèles des locuteurs (les locuteurs sont connus à l'avance par le système).

Le but du projet CoCAPs est de développer un dispositif qui peut fonctionner aussi dans des espaces où les locuteurs ne sont pas connus à l'avance par le système comme le cas des sales de réunion.

En effet, il faut utiliser un système qui peut fonctionner sans la présence des modèles des locuteurs, tel que la segmentation en locuteurs.

### 2.3. Segmentation en locuteurs

La segmentation en locuteurs [4] se base sur l'hypothèse qu'aucune information n'est connue à l'avance par le système, elle permet de découper le flux audio en segments ; en précisant le début, la fin ainsi que l'étiquette du locuteur auquel il correspond.

Pour résoudre le problème de la segmentation en locuteurs, on peut utiliser des techniques existantes en reconnaissance automatique du locuteur, telles que la paramétrisation de parole et la modélisation statistique du locuteur.

L'architecture des systèmes de segmentation en locuteurs est présentée dans la figure 4, elle se compose de plusieurs étapes :

1. Paramétrisation acoustique du signal de parole ;
2. Pré-segmentation acoustique, permet de détecter les segments contenant uniquement de la parole et, d'éliminer les segments qui contiennent du silence, de la musique et du bruit, etc ;
3. Détection de changement de locuteur, vise à obtenir des segments contenant de la parole appartenant à un seul locuteur ;
4. Regroupement des segments, permet de regrouper les segments trouvés pendant la phase précédente selon le locuteur où chaque groupe est identifié par l'étiquette du locuteur.

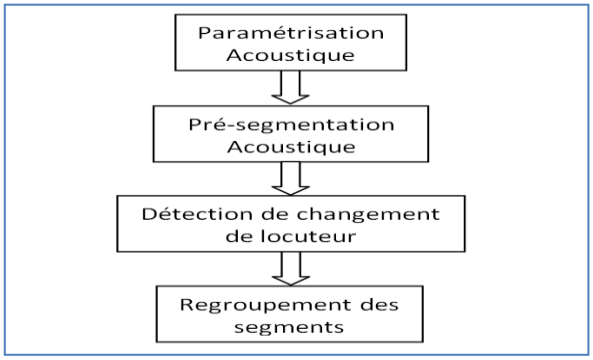


Figure 4 : Système de segmentation en locuteurs.

### 2.4. Reconnaissance des sons

La classification des sons est basée sur le modèle de mélanges de distribution de Gauss(*GMM*)[1]. Chacune des classes des sons de la vie courante est modélisée par un GMM. Elle 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.

### 2.5. Reconnaissance des expressions

Le système de reconnaissance des expressions est présenté dans la figure 5, il permet d'analyser la voix humaine captée par un microphone sous forme d'un signal audio pour la transcrire sous forme d'un texte exploitable par la machine en se basant sur les modèles acoustiques, les grammaires de mots et les modèles de langage. Les modèles de Markov cachés (*HMM*)[5] sont devenus la solution de référence pour tout dispositif de reconnaissance automatique de la parole.

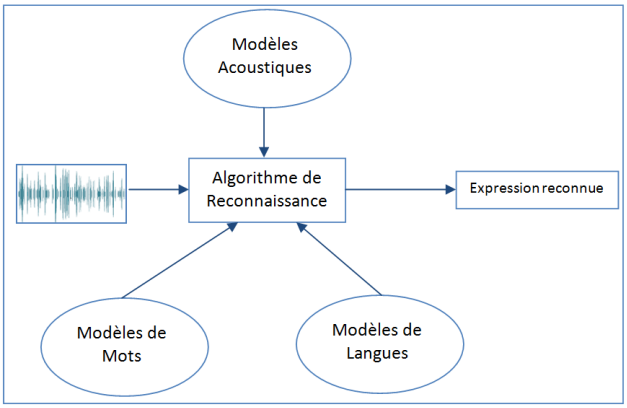


Figure 5 : Système de reconnaissance des expressions.

L'algorithme de reconnaissance de l'expression (phonème, mot ou phrase) se base sur un automate probabiliste (*HMM*), où on peut avoir dans chaque état une distribution statistique qui est un mélange de distributions de Gauss à covariance diagonale, ce qui permet d'obtenir une probabilité pour chaque vecteur observé.

Chaque mot présente une distribution de probabilité différente et le modèle HMM d'une phrase est la concaténation des modèles HMM formés pour les mots séparés.

### 2.6. Paramétrisation

La paramétrisation [4] ou le calcul des paramètres acoustiques à partir du signal de parole, consiste à transformer le signal échantillonné du son brut en paramètres acoustiques, dont le but est d'extraire l'information pertinente pour la tâche proposée. Parmi les paramètres utilisés on peut citer trois catégories principales :

- les paramètres spectraux : transformée de Fourier, bancs de filtres;
- les paramètres temporels : le taux de passage par zéro, le débit d'élocution;

- les paramètres cepstraux : LPCC *(Linear Prediction Cepstral Coefficient)*, LFCC *(Linear Frequency Cepstral Coefficient)*, MFCC (*Mel Frequency Cepstral Coefficient*).

Les paramètres cepstraux sont utilisés pour séparer l'influence de la source d'excitation vocale et celle du conduit vocal, cette dernière étant généralement la seule à être utilisée dans le domaine de la reconnaissance du locuteur. En effet, les paramètres acoustiques les plus utilisés pour la caractérisation du locuteur sont les coefficients cepstraux, notamment les coefficients MFCC.

Les coefficients MFCC [4] d'une trame de parole sont calculés de la façon suivante :

1. Après le filtrage de pré-accentuation, le signal de parole est d'abord découpé en fenêtres de taille fixe reparties de façon uniforme le long du signal.
2. La FFT (Fast Fourier Transform) de la trame est calculée. Ensuite, l'énergie est calculée en élevant au carré la valeur de la FFT. L'énergie est passée ensuite à travers chaque filtre Mel. Soit $S_k$ l'énergie du signal à la sortie du filtre *K*, nous avons maintenant $m_p$ (le nombre de filtres) paramètres $S_k$.
3. Le logarithme de $S_k$ est calculé.
4. Finalement les coefficients sont calculés en utilisant la DCT (Discrete Cosine Transform).

$$C_i = \sqrt{\frac{2}{m_p}} \left\{ \sum_{k=1}^{m_p} lo(S_k) \cos\left[ i \left( k - \frac{1}{2} \right) \frac{\pi}{m_p} \right] \right\} \quad Pour\ i = 1 \dots N$$

Où *N* est le nombre de coefficients MFCC que l'on souhaite obtenir.

2.7. La modélisation du locuteur

Les modèles utilisés pour la caractérisation du locuteur sont issus des modèles statistiques employés en reconnaissance de formes. Les modèles les plus utilisés dans la reconnaissance du locuteur sont les mixtures de gaussiennes GMM (Gaussian Mixture Models) et les modèles de Markov cachés HMM (Hidden Markov Models) qui sont largement employés en reconnaissance de la parole.

Ces modèles sont généralement utilisés en reconnaissance du locuteur dans des conditions dépendantes du texte (HMM) ainsi que dans des conditions indépendantes du texte (GMM) [6].

Cependant, Il existe bien d'autres modèles comme ceux issus de l'analyse prédictive, les réseaux de neurones ou les machines à support de vecteurs (SVM).

*2.7.1. Les modèles de mélange gaussiens*

Les GMMs [4] modélisent les séquences de vecteurs acoustiques, correspondant à un locuteur, par une somme pondérée de distributions gaussiennes multidimensionnelles.

Une distribution gaussienne multidimensionnelle est définie par la formule suivante :

$$P(x) = \frac{1}{\sqrt{(2\pi)^d |\sum|}} exp\left( -\frac{1}{2} (x - \mu)^T \sum^{-1} (x - \mu \right)$$

Où *d :* est la dimension d'un vecteur de paramètres noté *x* ;
μ: le vecteur moyen ;
Σ : la matrice de covariance estimée à partir de données $X = \{x_1, x_2, \dots, x_N\}$ ; selon les formules suivantes :

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

$$\sum = (X - \bar{X})(X - \bar{X})^T$$

Un GMM est défini par un ensemble de *T* distributions gaussiennes et de *T* poids $w_i$ associés à chaque distribution $P_i(x)$. Le nombre *T* est appelé l'ordre du modèle.

$$P(x) = \sum_{i=1}^{T} w_i P_i(x)$$

L'utilisation de GMM pour la modélisation du locuteur est motivée par deux interprétations :

- chaque distribution dans un modèle GMM est capable de représenter la structure spectrale d'une large classe phonétique. Ces classes représentent des configurations du conduit vocal spécifiques au locuteur et donc utiles pour la modélisation du locuteur ;
- une mixture de distributions gaussiennes donne une représentation approximative de la distribution à long terme de vecteurs acoustiques provenant des énoncés du même locuteur.

*2.7.2. Les modèles de Markov cachés*

Les HMMs [4] sont issus de la théorie des automates probabilistes. Un automate probabiliste est défini par une structure composée d'états, de transitions et par un ensemble de distributions de probabilités sur chaque état.

Un HMM peut être défini comme un quadruplet :

- un ensemble d'états $s_i$ avec $\pi_i$ les probabilités initiales d'états ;
- un ensemble de probabilités de transitions entre les états $a_{i,j} = P\left(\frac{i}{j}\right)$ avec $\sum_j a_{i,j} = 1$ ;

- un alphabet de symboles (pas forcement fini) ;
- un ensemble de densités de probabilités (exemple : GMM) pour l'émission de symboles associés à chaque état $b_i(x)$.

Deux états supplémentaires, début et fin, peuvent être ajoutés pour permettre d'interconnecter des HMMs ou d'imposer des contraintes sur la séquence d'états.

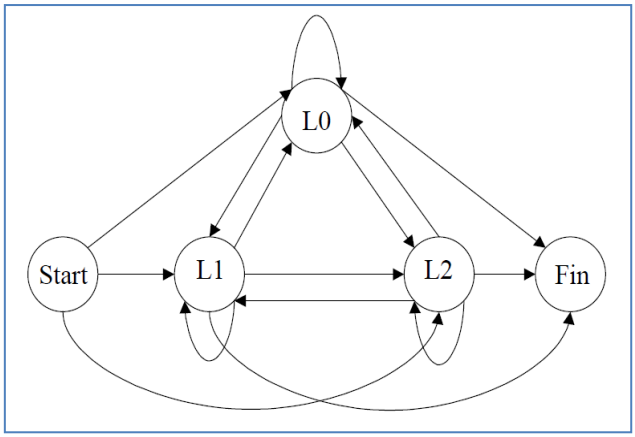


Figure 6 : Un modèle HMM ergodique (toutes les transitions sont possibles).

Les modèles de Markov cachés ont connu un succès important en reconnaissance de la parole où ils se sont imposés comme l'un des modèles de référence. Dans la reconnaissance du locuteur ils sont aussi utilisés surtout en reconnaissance dépendante du texte où il existe une forte connaissance à priori sur l'énoncé du test.

Les HMMs sont aussi utilisées pour résoudre le problème de segmentation c'est-à-dire de découpage d'une séquence en sous-séquences de différents types comme par exemple découpage d'un signal audio en zones contenant une certaine caractéristique (parole/silence).

## III. Mise en œuvre de L'application « Classification des sons »

L'application « classification des sons » est développée dans le cadre du stage d'Adrian NICOLICEA, elle vise à classifier les sons de la vie courante et à détecter des expressions de détresse. Dans notre projet CoCAPs on se base sur cette application afin de réaliser le deuxième objectif qui est la détection des situations de détresse en améliorant les résultats obtenus.

### 3.1. Principe de fonctionnement

L'architecture de l'application développée est présentée en figure 7, c'est une application de type client-serveur et elle a pour objectifs de détecter, classifier et enregistrer du son.

La classification se fait en deux étapes : la première consiste à diviser le flux audio entre la voix et le son de la vie courante, et la deuxième étape consiste à détecter des expressions de détresse.

Afin de résoudre le problème de détection de la détresse et de situations critiques, le son doit être continuellement analysé. Pour cela, une boucle infinie a été utilisée pour calculer et stocker le son détecté, l'étiquette, le rapport signal à bruit (SNR), le temps de détection (*Timestamp*) et la *log-vraisemblance* (LLK). Le stockage se fait dans une liste de type FIFO (First in – First out).

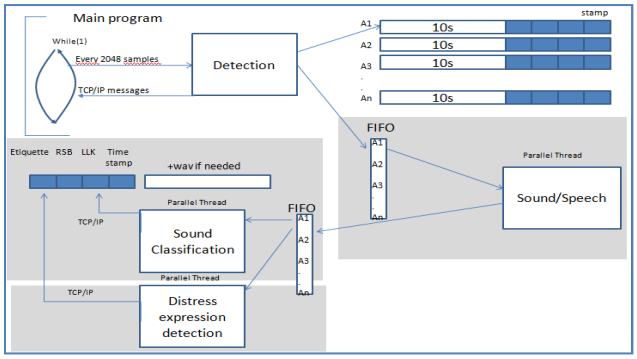


Figure 7 : L'architecture de l'application.

Pour assurer un fonctionnement en temps réel de l'application, des threads (fil d'exécution) parallèles pour les programmes de détection et de classification ont été utilisés.

La communication client-serveur se fait via le protocole TCP-IP où l'utilisateur « client » envoie le message « START » pour démarrer à distance le programme de classification et d'enregistrement au niveau du serveur. Ce dernier renvoie au client d'une façon continue les informations récupérées et calculées (étiquette, SNR, LLK, Timestamp, son détecté). Afin d'avoir un contrôle sur la fonctionnalité et la durée d'utilisation, le client peut arrêter l'enregistrement sur le serveur en envoyant le message « STOP ».

Le développement est réalisé sur la carte Raspberry Pi1 en utilisant la bibliothèque ALIZE développée par le Laboratoire d'Informatique d'Avignon (*LIA*) et mettant en œuvre d'une part les modèles GMM (*Gaussian Mixture Models*) pour la classification du son et d'autre part les modèles HMM (*Hidden Markov Model*) pour la reconnaissance de la parole, ceci afin de détecter des situations de détresse [3].

### 3.2. Premières évaluations

Afin d'évaluer notre application de détection de situations de détresse, 18 classes de son de la vie quotidienne ont été utilisées. Elles ont été sélectionnées par Mohamed SEHILI lors de sa thèse [2] pour répondre aux besoins identifiés pour le suivi des personnes âgées. Un de nos objectifs est de les faire évoluer par rapport aux besoins du projet CoCAps : les classes comprennent cette fois des sons humains (respiration, toux, pleurs,…) et des sons autres qu'humains (moteur, fenêtres cassées,...). La fréquence d'échantillonnage des fichiers audio utilisée est de 16 kHz.

Les résultats ci-dessous ont été réalisés dans le cadre du stage d'Adrian NICOLICEA [3], ils ont été obtenus en répétant chaque son 20 fois sur différentes tonalités.

Le pourcentage de reconnaissances correctes est aussi évalué en présentant différentes compilations de son au système de détection.

| | Cough | DoorClapping | Paper | Breathlessnes | Laugh | Sneeze | Dishes | Yawn | ElectricalShaver | HandsClapping | DiscardedSounds | Speech |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cough | 95% | 5% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | |
| DoorClapping | 10% | 70% | 20% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | |
| Paper | 0% | 20% | 70% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 10% | |
| Breathlessness | 10% | 0% | 10% | 0% | 0% | 0% | 0% | 5% | 5% | 60% | 10% | |
| Laugh | 90% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 10%(un docteur) |
| Sneeze | 5% | 90% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 5% | 0% | |
| Dishes | 0% | 0% | 10% | 0% | 0% | 0% | 45% | 0% | 0% | 45% | 0% | |
| Yawn | 0% | 0% | 0% | 0% | 10% | 10% | 10% | 10% | 10% | 0% | 50% | |
| ElectricalShaver | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 90% | 0% | 10% | |
| HandsClapping | 0% | 10% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 90% | 0% | |

Table1 : Classification du son (MFCC)

| | AidezMoi | Appelez | AuSecours | CaVaPas | Help | JaiBesoi | JeMeSens | UnDocteur | UneInfirmie | UnToubib | Vite | Jai | NormalSc | Sound |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AidezMoi | 40% | 0% | 0% | 0% | 0% | 10% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | |
| Appelez | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 100% | 0% | |
| AuSecours | 0% | 0% | 80% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 10% | 0% | 10%(cough) |
| CaVaPas | 10% | 20% | 0% | 0% | 0% | 30% | 10% | 0% | 0% | 0% | 30% | 0% | 0% | |
| Help | 0% | 0% | 0% | 0% | 80% | 0% | 0% | 10% | 0% | 0% | 10% | 0% | 0% | |
| JaiBesoin | 0% | 0% | 0% | 0% | 0% | 100% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | |
| JeMeSens | 0% | 0% | 0% | 0% | 0% | 100% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | |
| UnDocteur | 20% | 0% | 0% | 0% | 0% | 10% | 0% | 20% | 0% | 0% | 0% | 0% | 0% | 20%(cough) |
| UneInfirmiere | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 20% | 0% | | |
| UnToubib | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 50% | 0% | 50% | |
| Vite | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 80% | 0% | 0% | 10%yawn, 10%sneeze,20%cough |

Table 2 : Classification de la voix (MFCC).

| | Cough | DoorClapping | Paper | Breathlessnes | Laugh | Sneeze | Dishes | Yawn | ElectricalShaver | HandsClapping | DiscardedSounds |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Cough | 85% | 15% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% |
| DoorClapping | 40% | 60% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% |
| Paper | 0% | 50% | 50% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% |
| Breathlessness | 100% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% |
| Laugh | 20% | 0% | 0% | 0% | 60% | 0% | 0% | 0% | 20% | 0% | 0% |
| Sneeze | 0% | 30% | 10% | 0% | 0% | 10% | 0% | 0% | 0% | 50% | 0% |
| Dishes | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 100% | 0% |
| Yawn | 30% | 30% | 0% | 0% | 0% | 0% | 0% | 0% | 10% | 30% | 0% |
| ElectricalShaver | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 100% | 0% | 0% |
| HandsClapping | 0% | 30% | 0% | 0% | 0% | 0% | 0% | 0% | 10% | 60% | 0% |

Table 3 : Classification du son (LFCC).

| | AidezMoi | Appelez | Vite | A l'aide |
|---|---|---|---|---|
| AidezMoi | 50% | | | |
| Appelez | | 60% | | |
| Vite | | | 50% | |
| A l'aide | | | | 80% |

Table 4 : Reconnaissance de la parole (Situations de détresse).

Après les premières évaluations de l'application, on peut constater que l'utilisation du coefficient MFCC donnera de meilleurs résultats au niveau de la classification du son comparés (Table 1) à ceux obtenus avec l'utilisation du coefficient LFCC (Table3).

Pour la classification de la voix (Table 2), on peut remarquer que les résultats sont meilleurs (80%-100%) au niveau de quelques expressions de détresse telles que, Aidez-Moi, J'ai besoin, et Au secours.

Par contre les premiers résultats de l'évaluation du système de la reconnaissance de la parole (Table 4) appliqué sur des expressions de détresse sont moyens et des améliorations sont envisagées pour obtenir des résultats plus affinés.

## IV. Conclusion et Perspectives

Dans cet article nous avons présenté nos travaux de recherche sur la partie d'analyse du son du projet CoCAPs.

Le développement du système de segmentation en locuteurs est en cours afin de réaliser le premier objectif du projet qui est la détection du nombre de personnes.

Dans un deuxième temps l'utilisation de la reconnaissance de parole est envisagée pour répondre au deuxième objectif du projet CoCAPs qui est la détection des situations de détresse.Ainsi cela permettra une amélioration de l'application présentée ci-dessus.


## REMERCIEMENTS

Les auteurs tiennent à remercier BPI France, les Conseils Régionaux du Limousin et de Rhône-Alpes associé au programme FEDER, le conseil départemental de l'Isère, et la communauté d'agglomération Bourges Plus, pour leur soutien financier au projet CoCAPs. Le projet CoCAPs, issu du FUI N°20, est également soutenu par les pôles de compétitivité S2E2 et Minalogic.


## Bibliographie


[1] D. ISTRATE, "Détection Et Reconnaissance Des Sons Pour La Surveillance Médicale ", thèse de doctorat, Laboratoire CLIPS-IMAG, INPG, 2003.

[2] M. A. SEHILI, "Reconnaissance Des Sons De L'environnement Dans Un Contexte Domotique ", thèse de doctorat, ESIGETEL - Université d'Evry Val D'Essonne 2013.

[3] A. NICOLICEA, "Sound Recognition For Medical Remote Monitoring", Rapport de stage M2, ESME Sudria - Université de Craiova, 2014.

[4] D. MORARU, "Segmentation en locuteurs de documents audios et audiovisuels: application à la recherche d'information multimédia", these de doctorat, Institut National Polytechnique de Grenoble, 2004.

[5] L.R. Rabiner, « A tutorial on hidden Markov models and selected applications in speech recognition", Proceedings of the IEEE, Vol. 77, Issue: 2, Feb 1989, pp 257-286, DOI: 10.1109/5.18626

[6] Meignier, S., Bonastre, J., Fredouille, C., and Merlin, T. (2000). Evolutive HMM for multi-speaker tracking system. International Conference on Audio, Speech and signal processing ICASSP '2000, pp. 543–547, Istanbul, République Turque.

[7] Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. Digital signal processing, 10(1-3), 19-41.