الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne démocratique et populaire

وزارة التعليم العالي والبحث العلمي
Ministère de l'enseignement supérieur et de la recherche scientifique

جامعة سعد دحلب البليدة
Université SAAD DAHLAB de BLIDA

كلية التكنولوجيا
Faculté de Technologie

قسم الإلكترونيك
Département d'Électronique

Mémoire de Projet de Fin d'Études

Présenté par

Yasmine BENGRAB

Khadidja HADJALA

Pour l'obtention du diplôme de Master en Électronique

Spécialité : Instrumentation

Thème

Reconnaissance des évènements audio en vue de la mise en œuvre d'un système de surveillance audio

Proposé par :

Fayçal YKHLEF, Maître de Recherche B, CDTA, Alger.

Farid YKHLEF, Maître de Conférences A, Université SAAD DAHLEB, BLIDA.

Année Universitaire 2019-2020

REMERCIEMENTS

Tout d’abord, nous remercions ALLAH Sobhanou de nous avoir accordé la volonté et le courage d'entreprendre et d'achever ce travail.

D’emblée, nous devons une reconnaissance personnelle très profonde à notre promoteur et encadreur Dr. Fayçal YKHLEF, Maitre de Recherche B au CDTA, pour avoir orienté et enrichi notre travail. Nous le remercions pour sa disponibilité, ses précieux conseils, sa confiance malgré nos connaissances plutôt légères dans le domaine de traitement de signal. Nous le remercions aussi pour son souci du détail qui a abouti à la réalisation de ce mémoire.

Nous remercions tout le personnel et les chercheurs du Centre de Développement des Technologies Avancées (C.D.T.A), et en particulier ceux de la Division Architecture des Systèmes et Multimédias (ASM) pour leur accueil et leur soutien.

Nous tenons à remercier notre co-encadreur Dr. Farid YKHLEF, Maitre de Conférence A au niveau de l’Université SAAD DAHLEB, pour sa disponibilité ses conseils et son sens d’écoute et d’échange.

Nos remerciements vont également aux membres du jury pour avoir accepté d’examiner notre travail et de l’enrichir par leurs propositions.

Nous souhaitons aussi adresser nos remerciements au corps professoral et administratif de l’Université de BLIDA qui a contribué à la réussite de nos études universitaires.

DEDICACE

Du profond de mon cœur, je remercie Allah le tout puissant.

À Ma chère mère, que nulle dédicace ne puisse exprimer mes sincères sentiments pour sa patience illimitée, son encouragement continu, son aide, en témoignage de mon profond amour et respect pour ses sacrifices. Que dieu le tout puissant te protège et te garde à mes côtés.

À Mon père, qui a toujours prié pour moi, ma soutenu et ma épauler pour que je puisse atteindre mes objectifs. Que dieu te protège et te garde pour moi. Que ce travail soit le témoignage de ma gratitude et de mon affection.

À Mon frère Fiçal, qui m'a toujours poussé et motivé dans mes études. Merci énormément pour ton soutien plus que précieux, merci pour ton grand cœur et ta qualité qui serait trop longues à énumérer. Je t'aime de tout mon cœur.

À Mes sœurs Lamia, Imene et Ikram. A ma belle-sœur Imene. Que dieu illumine leur voie de succès et de réussite. Ma vie ne serait pas aussi magique sans votre présence et votre amour.

C'est un moment de plaisir de remercier ma très chère amie Khadidja HADJALA, mon binôme. J'ai du mal à trouver les mots d'amour, de reconnaissance et de gratitude.

À mon très cher neveux Wassim, et mes adorables nièces Lina et Inès. À tous mes amis que j'aime.

Yasmine

DEDICACE

C’est avec un grand plaisir que je dédie ce travail :

À mes parents :

À ma très chère maman Djamila. Au meilleur des pères Mohamed, qui n’ont jamais cessé de me soutenir et de m’encourager pour que je puisse réaliser mes objectifs. Aucune dédicace ne saurait exprimer mon respect, mon amour éternel et ma considération pour les sacrifices que vous avez faits pour mon instruction et mon bien être. Je tiens à vous remercier pour tout le soutien et l'amour que vous me portez depuis mon enfance. Que cette humble œuvre comble vos désirs et qu’elle soit le fruit de vos innombrables sacrifices. Que Allah vous accorde santé, bonheur et longue vie et faire en sorte que jamais je ne vous déçoive.

À mes chers frères Abdelhak et Walid et ma chère sœur Wafa et ma belle-sœur Asma, en témoignage de mon affection fraternelle, de ma profonde tendresse et reconnaissance, je vous souhaite une vie pleine de bonheur et un avenir radieux plein de réussite.

À ma petite nièce chérie Insaf, a tous mes oncles, tantes, mes cousins et cousines. A toute ma famille. Merci pour tout le soutien que vous m’avez apporté.

Je tiens à remercier spécialement ma chère amie Yasmine BENGRAB, qui a été mon binôme dans ce mémoire, pour sa patience et sa persévérance. Je souhaite que l'amitié que nous a réunie persiste pour toujours et que nous arrivons à réaliser nos rêves.

Je voudrais exprimer ma reconnaissance envers tous mes chers et meilleurs amis sans exception qui m'ont apporté leur support moral et intellectuel tout au long de mon cursus.

A tous ceux qui ont était là pour moi,

À tous ceux qui me sont chers,

A mes collègues de promotion INS 2020.

Khadidja

RESUME

ملخص:

الغرض الرئيسي من هذا المشروع هو تصميم جهاز التعرف على الصوت البيئي باستخدام عينات صغيرة الحجم. أداة التعرف هي إحدى الوحدات النمطية الهامة الرئيسية المدمجة في أنظمة المراقبة الصوتية. نركز في هذا المشروع على التعرف على ثلاثة أحداث صوتية رئيسية: (1) صراخ الإنسان ، (2) أجهزة إنذار السيارة ، (3) الزجاج المكسور. نقترح مخطط تصنيف يعتمد على التشوه الزمني الديناميكي. علاوة على ذلك ، نقدم منهجية تتمثل في استخدام دفتر رموز يتكون من عدة بيانات مرجعية لحساب المسافات بين تسلسلات الاختبار. قارنا أداء التعرف على الطريقة المقترحة باستخدام طريقتين مختلفتين لاستخراج الميزات: MFCCs (معاملات سيبسترال ذات تردد الميل) و LPCs (معاملات التنبؤ الخطية). وجدنا أن الطريقة المقترحة حققت دقة 83.33٪ باستخدام 5 سمات LPC فقط. علاوة على ذلك ، تم الحصول على هذه النتيجة باستخدام مقطع صوتي يبلغ 0.2 ثانية فقط. غير أن، بالنسبة إلى MFCCs ، تم تحقيق نفس الدقة عندما (1) كانت مدة مقاطع الصوت 0.6 ثانية و (2) كان عدد السمات 30 (بما في ذلك مشتقات MFCCs). نستنتج أن LPCs أكثر ملاءمة من MFCCs..

الكلمات الدالة: التعرف على الأحداث الصوتية ، معالجة الإشارات الصوتية ، MFCC، LPC ، التشوه الزمني الديناميكي.

Résumé :

Ce projet consiste à concevoir une méthode de reconnaissance des sons de l'environnement en utilisant peu de données. Il s'agit d'un module important pour la conception des systèmes de surveillance audio. Nous nous sommes intéressés à la reconnaissance de trois catégories de sons : (i) cris humains, (ii) alarmes de voitures et (iii) bris de glace. Le schéma de classification que nous proposons est basé sur la déformation temporelle dynamique. Nous appliquons une méthodologie qui consiste à utiliser une multitude de données de références (codebook) pour le calcul des distances entre les séquences. Nous comparons les performances de la méthode de reconnaissance en utilisant deux techniques d'extraction d'attributs : les MFCCs (Mel-Frequency Cepstral Coefficients) et les LPCs (Linear Prediction Coefficients). Nous avons trouvé que la méthode proposée a atteint un taux de reconnaissance de **83.33%** en utilisant seulement 5 attributs LPCs. De plus, ce résultat a été obtenu en exploitant seulement une durée égale à 0.2s du segment sonore.

Cependant, pour les MFCCs, le même taux de reconnaissance a été atteint en utilisant (i) une durée du segment sonore de 0.6s et (ii) 30 attributs (incluant les dérivées des MFCCs). Nous concluons que les LPCs sont plus appropriés que les MFCCs.

Mots clés : Reconnaissance des évènements acoustiques, traitement du signal audio, MFCC, LPC, la déformation temporelle dynamique.

Abstract :

The main purpose of this project is to design an environmental sound recognizer using small size samples. The recognizer is one of the main important modules that are incorporated into audio surveillance systems. We are focusing in this project on the recognition of three main sound events: (i) human screams, (ii) car alarms and (iii) broken glasses. We propose a classification scheme based on dynamic time warping. Furthermore, we introduce a methodology which consists in using a codebook composed of several templates to compute distances between test sequences. We compare the recognition performance of the proposed method using two different feature extraction techniques: MFCCs (Mel-Frequency Cepstral Coefficients) and LPCs (Linear Prediction Coefficients). We found that the proposed method achieved an accuracy of 83.33% using only 5 LPC attributes. Moreover, this result was obtained using a sound segment of only 0.2s. However, for the MFCCs, the same accuracy was achieved when the (i) duration of sound segments was 0.6s and (ii) the number of attributes was 30 (including derivatives of the MFCCs). We conclude that LPCs are more appropriate than MFCCs.

Keywords: Sound events recognition, audio processing, MFCC, LPC, DTW.

LISTE DES ACRONYMES ET ABREVIATIONS

ACC: *Accuracy*

DCT: *Discrete Cosine Transform*

DTW: *Dynamic Time Warping*

ED: *Euclidean Distance*

ER: *Error Rate*

FFT: *Fast Fourier Transform*

FN: *False Negative*

FP: *False Positive*

LPCs: *Linear prediction coefficients*

MFCCs: *Mel-Frequency Cepstral Coefficients*

SVMs: *Support Vector Machines*

SER: *Sound Events Recognition*

TP: *True Positive*

TN: *True Negative*

TPR: *True Positive Rate*

TABLE DES MATIERES

LISTE DES FIGURES

LISTE DES TABLEAUX

Chapitre 1 : Introduction

1.1 Motivations

De nos jours, les systèmes de surveillance jouent un rôle important dans la sécurité publique et routière. Les systèmes de surveillance de troisième génération utilisent une multitude de capteurs supplémentaires aux caméras pour assurer un contrôle automatique. On peut citer, les capteurs infrarouges, les capteurs thermiques, les microphones et les capteurs de mouvements.

Particulièrement, les microphones peuvent capturer des événements audio qui sont en dehors du champ de vision des caméras et permettent de diriger automatiquement les cameras vers les sources du danger (coups de feu, bris de glace, accidents de voitures). L'accomplissement de cette tâche nécessite trois phases importantes : (i) la détection des évènements audio, (i) la reconnaissance des sons, et (iii) la localisation de l'endroit exacte de la source sonore. Nous nous focalisons dans ce mémoire sur la reconnaissance.

Le CDTA est en train de conduire un projet socioéconomique sur la conception et l'implémentation de méthodes intelligentes pour la gestion et la surveillance des parkings. L'incorporation de l'information audio pour la surveillance des lieux est un des objectifs de ce projet.

Notre stage au CDTA consiste à concevoir et implémenter le module de reconnaissance des sons de l'environnement. Les exigences ainsi que les objectifs de cette tâche sont résumées comme suit :

- La reconnaissance des sons de l'environnement doit utiliser des méthodes à complexité réduite pour faciliter son implémentation en temps réel,
- Ce module de reconnaissance doit être adapté à une plateforme de détection des sons impulsifs qui existe au niveau du laboratoire ASM,
- Le début exact du son est considéré connu,
- La durée du son nécessaire à la tâche de reconnaissance doit être courte,

1.2 Contributions

Nous nous focalisons dans le cadre de ce projet à la reconnaissance de certains sons de l'environnement, à savoir : les cris humains, les bris de glace, et les alarmes de voitures.

Le schéma de classification que nous proposons est basé sur la déformation temporelle dynamique. Notre choix est motivé par le fait que nous construisons un modèle de reconnaissance en utilisant peu de données. Nous avons utilisé une multitude de données de références (codebook) pour le calcul des distances temporelles entre les séquences. Notre objectif est de comparer les performances de la reconnaissance en utilisant deux techniques d'extraction d'attributs : les MFCCs (Mel-Frequency Cepstral Coefficients) et les LPCs (Linear Prediction Coefficients). Les MFCCs sont basés sur le processus de la perception auditive humaine tandis que les LPCs sont estimées à partir du modèle phonatoire de production de la parole (le modèle source filtre). Nous avons construit un corpus sonore qui inclut les sons cités ci-dessus pour valider notre comparaison.

1.3 Impact du projet

Le travail accompli dans ce projet est très utile pour la sécurité des citoyens. Les technologies de surveillance à base des informations audio ne sont pas très répondues en Algérie. A cet effet, la maitrise de ces technologies peut aider à développer le secteur socioéconomique de notre pays.

Les technologies de surveillance audio ainsi que les autres technologies développées au laboratoire ASM du CDTA peuvent intéresser plusieurs institutions en Algérie. Nous citons : la sureté interne, les militaires, les banques, les entreprises, les universités, et les écoles.

1.4 Organisation du mémoire

Le mémoire est organisé comme suit :

Le deuxième chapitre : présente des généralités sur les systèmes de surveillance et se focalise principalement sur l'apport de la modalité audio pour la surveillance des lieux.

Le troisième chapitre : présente deux points essentiels : (i) la méthode de reconnaissance des évènements sonores, et (ii) les métriques utilisées pour l'évaluer.

Le quatrième chapitre : expose les résultats expérimentaux ainsi que leur interprétation.

Le cinquième chapitre : comprend une conclusion générale et des perspectives (travaux futures).

Chapitre 2 : Généralités sur les systèmes de surveillance

2.1 Introduction

Dans ce chapitre nous présentons des généralités sur les systèmes de surveillance. Nous abordons dans la première phase quatre points essentiels : (i) définition, (ii) historique, (iii) applications, et (iv) évolution des systèmes de surveillance. Dans la deuxième phase, nous nous focalisons sur l'apport de la modalité audio pour la surveillance. Nous terminons ce chapitre par une conclusion.

2.2 Systèmes de surveillance

2.2.1 Définition

Les systèmes de surveillance sont des outils technologiques importants pour la détection, la surveillance et le suivi des activités malveillantes dans un certain environnement.

Particulièrement, la surveillance visuelle vise à comprendre et décrire le comportement des objets dans une scène donnée [1] [2] [3]. Avec les avancés de la technologie sans fil et de l'intelligence artificielle, plusieurs types de capteurs sont employés pour surveiller collectivement les régions d'intérêt. Ces capteurs sont capables de capturer et traiter des images, des vidéos ainsi que des signaux, et d'envoyer à la station centrale seulement la donnée nécessaire à l'interprétation de l'activité [1].

Les systèmes de surveillance sont des outils importants pour aider les humains à étendre leurs capacités de perception et de raisonnement sur diverses situations d'intérêt.

2.2.2 Historique

Le premier système de vidéosurveillance a été conçu en 1942 par la société Allemande Siemens AG. Il s'agit particulièrement d'un système de vidéosurveillance qui permet d'observer les différentes étapes de lancement des fusées V2 en vue de détecter des éventuels dysfonctionnements ou des comportements errant dans le processus de lancement [4].

Vingt-sept ans après, à New York, M. V. B.Brown [5] et al. ont proposé un système de sécurité à domicile qui comprend des judas, des caméras, des microphones, des moniteurs et des télécommandes pour permettre à quelqu'un dans sa maison de vérifier qui était à la porte d'entrée et de répondre en conséquence (Figure 2.1). Ce système a ouvert la voie aux futurs inventeurs et aux systèmes de sécurité modernes.

En 1970, le Royaume-Uni a installé pour la première fois des caméras de surveillances dans les endroits public afin de détecter et enregistrer les événements anormaux [6].

Dans les années 90, les caméras de surveillance ont été utilisées dans plusieurs endroits. À titre d'exemple, on peut citer : les routes, les transports, les parkings, et les administrations [6].

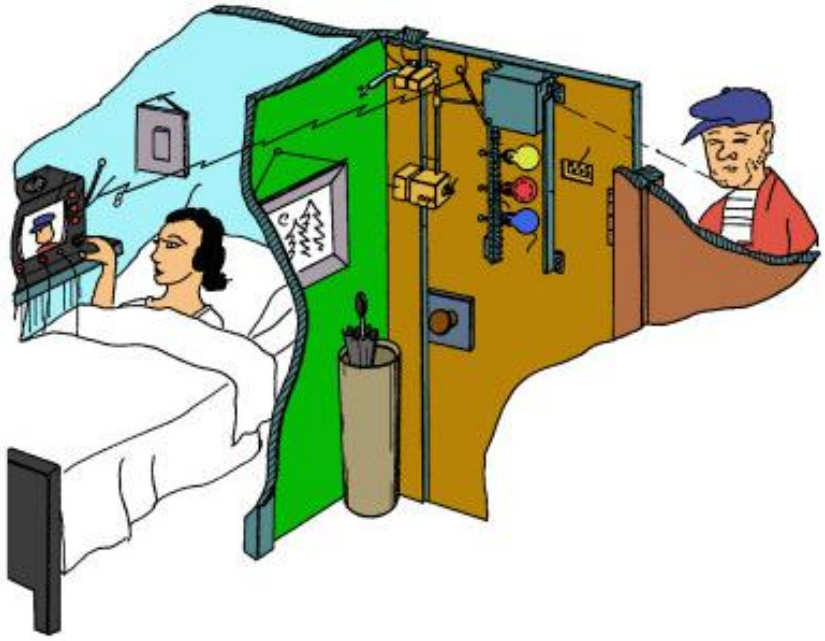

Figure 2.1: Système de surveillance proposé par Van Brown [5].

2.2.3 Applications

Dans un monde où les taux de criminalité et des attaques terroristes sont très élevés, les systèmes de surveillance sont devenus une solution inévitable. Un bref résumé sur ces applications est donné ci-dessous.

- Surveillance de la circulation sur les routes : cela comprend, la mesure de la vitesse du véhicule, la détection du passage au feu rouge, l'identification des voies inutiles et les autres infractions du code de la route [1] [7].
- Sécurité publique et commerciale : il s'agit de surveiller les lieux publics pour la détection des accidents et la prévention des crimes [1] [8]. Cette application inclue la surveillance des écoles,

les supermarchés, les théâtres, les grands magasins, les parkings, les stades, les aéroports, les chemins de fer, les métros et les lieux maritimes.

- Surveillance et étude environnementale : Cela couvre la surveillance des incendies de forêt, la pollution, les habitats, les animaux, les maladies des plantes et l'océanographique. La surveillance et la préservation des sites historiques sont aussi incluses dans cette catégorie [1] [9]**.**
- Applications militaires : On peut citer : les patrouilles des frontières, la mesure du flux des réfugiés, la surveillance des régions autour des bases militaires et l'aide au commandement des champs de bataille [10].
- Contrôle de qualité : Cela comprend la surveillance industrielle et automobile ainsi que les sites de production [1] [11]**.**
- Chambres intelligentes et sécurité personnelle : cela consiste à fournir une assistance médicale en observant les activités de personnes âgées et l'efficacité des traitements médicaux [12]. La détection des événements anormaux dans les maisons, tels que les vols, est incluse dans cette catégorie.

2.2.4 Evolution des systèmes de surveillance

D'un point de vue technologique, les systèmes de surveillance sont classés en quatre générations comme le montre la Figure 2.2.

Ces systèmes ont évolué d'une simple surveillance contrôlée par un opérateur, à une surveillance intelligente, automatisée et intégrée. Chaque génération s'appuie sur la précédente et apporte des changements en termes de plateformes ou types d'algorithmes nécessaires à aborder un large spectre de domaines de recherche et d'applications. L'intérêt pour la surveillance des lieux a connu une croissance remarquable en raison des problèmes de sécurité et de sûreté, en particulier après les évènements du 9/11 [1].

a) Système de surveillance de première génération

Les systèmes de surveillance de première génération sont des systèmes centrés sur l'humain (l'opérateur). Ils sont seulement utilisés pour la visualisation des lieux. Les flux vidéo capturés sous formes de signaux analogiques sont simplement transmis à une salle de contrôle à distance et affichés sur de grands moniteurs. La mission de l'opérateur est d'analyser, interpréter et classer les observations.

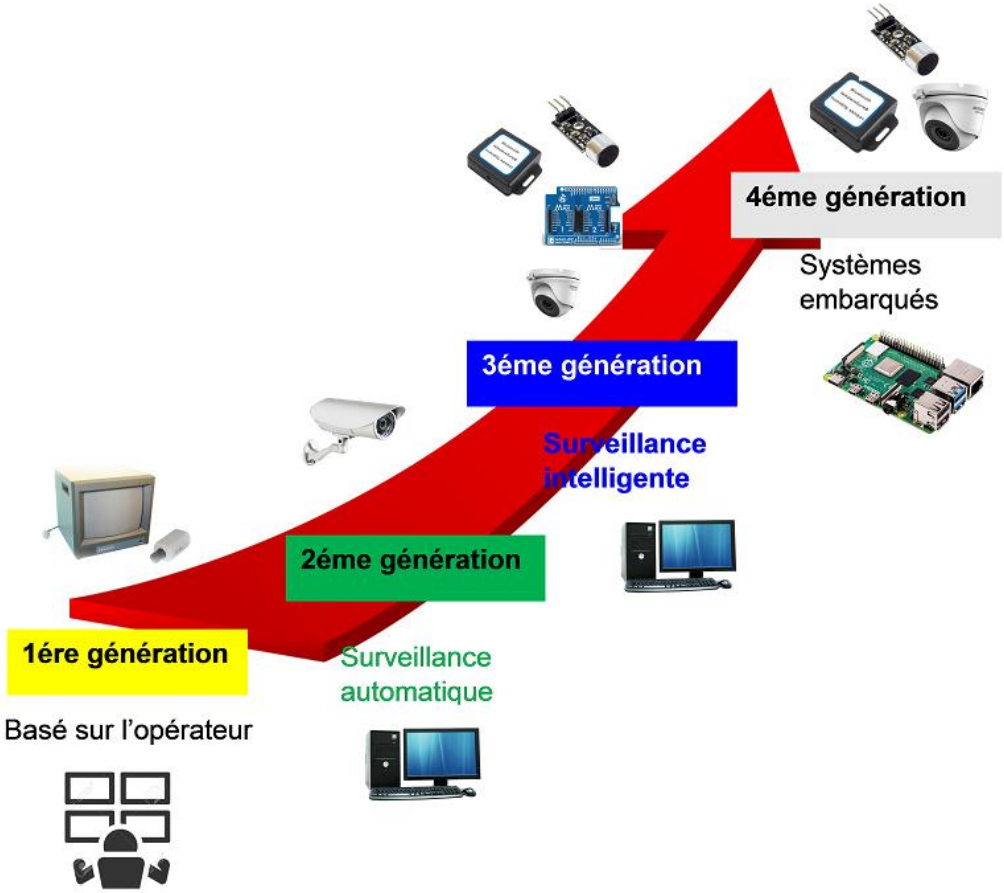

Figure 2.2: Evolution des systèmes de surveillance

Ces systèmes ne peuvent pas garantir une surveillance rigoureuse, longue et stable [13]. Avec l'augmentation du nombre de caméras, les tâches assignées aux opérateurs deviennent de plus en plus difficiles.

b) Système de surveillance de deuxième génération

Pour fournir une meilleure qualité de surveillance, la deuxième génération a été développée. Dans cette génération, les techniques de vision par ordinateur pour la détection et le suivi d'objets et l'analyse des scènes ont été adoptées en vue d'aider l'opérateur et focaliser l'attention particulièrement sur les situations anormales [1]. Cette génération a commencé avec l'introduction des caméras réseau (Caméra IP). La vidéo capturée est envoyée via des commutateurs réseau et affichée sur un PC. Ces systèmes sont entièrement numériques. Par rapport à la génération précédente, la qualité d'image a été considérablement améliorée.

L'inconvénient majeur de ces systèmes est que tous les traitements sont centralisés dans une station centrale (Ordinateur de contrôle).

c) Système de surveillance de troisième génération

Avec les progrès des réseaux informatiques mobiles et fixes, les systèmes de surveillance de troisième génération ont été développés. Le traitement numérique de l'information est distribué à divers niveaux du réseau. L'acquisition de l'information est achevée en utilisant une multitude de capteurs pour reconnaître les informations issues des différentes modalités (visuelle, infrarouge, thermique, audio, mouvement etc.) [14]. Le traitement est décalé de la station centrale vers des capteurs équipés de processeurs plus intelligents capables d'effectuer des tâches sur site.

En effectuant un certain nombre de traitement au niveau des caméras, le système transmet des connaissances, plutôt que des pixels, à la station centrale. La même opération est effectuée au niveau des autres capteurs (audio, mouvement, température, etc.). À ce stade et avec la disponibilité de la puissance de calcul, les recherches ont été orientés vers le développement de techniques de traitement distribué en temps réel [1].

En particulier, les progrès réalisés dans la conception de réseaux à large bande passante ont rendu ces systèmes utiles pour des applications de circulation dans les routes, la sécurité publique et la sécurité commerciale.

D'un point de vue pratique, la surveillance des zones non habitées, les forêts, et les montagnes n'est pas réalisable avec les systèmes de troisième génération principalement en raison de leurs caractéristiques intrinsèques. On peut résumer ces caractéristiques comme suit :

- Volume important,
- Consommation d'énergie élevée,
- Dépendance sur une station centrale (ordinateur).

d) Système de surveillance de quatrième génération

Pour faire face aux défaillances des systèmes de troisième génération, les systèmes de surveillance de quatrième génération ont vu le jour. Une évolution considérable vers les plates-formes embarquées a eu lieu vu leur adaptabilité aux environnements non habitées, aux forêts, et aux montagnes.

Avec les avancés de la fabrication de circuits intégrés, il devient de plus en plus possible de développer des capteurs et des systèmes distribués à faible puissance qui sont plus adéquat aux régions non habitées et aux situations d'urgence. De plus, ces systèmes peuvent assurer une surveillance plus appropriée et un traitement plus exact sans la nécessité d'une station de calcul centrale (Ordinateur) [1].

2.3 Modalité audio pour la surveillance

L'utilisation de la modalité audio pour la surveillance a été introduite par les systèmes de surveillance de la troisième génération. Cette modalité permet de capturer des événements audio qui sont en dehors du champ de vision de la caméra (Figure 2.3) [6]. L'utilisation de cette modalité pour la surveillance est très nécessaire lorsque les conditions climatiques sont défavorables.

Figure 2.3: Apport de la modalité audio pour la surveillance

Une fois la détection, la reconnaissance et la localisation de la position exacte de l'événement audio a été achevée, la caméra tourne vers la personne suspecté et fournie une vision complète sur la scène captée.

2.3.1 Evènements audio

La nature acoustique des événements anormaux est impulsive [15] . Ce type de sons peut surgir dans différentes situations de notre vie. Parmi ces sons, on peut citer : les cris humains, les explosions des pneus, les bris de glace, les coups de feu, et les aboiements des chiens. Un son impulsif correspond à la réponse vibratoire d'un choc bref. D'un point de vue perceptif, de nombreux tests psychoacoustiques montrent que l'oreille humaine est capable d'identifier et reconnaitre un son impulsif à partir de son amortissement et son spectre [16]. L'objectif des méthodes développées dans notre travail est la reconnaissance automatique de ces sons.

2.3.2 Détection, reconnaissance et localisation des événements audio

L'opération de détection fait référence à l'identification d'un événement dans un signal audio. Il s'agit de la discrimination entre les sons impulsifs et les autres sons de l'environnement. La seule information fournie par cette opération est qu'un événement anormal s'est produit suite à une augmentation soudaine de l'énergie [17].

L'opération de reconnaissance (du son impulsif) consiste à connaitre le type exact du son produit (Figure 2.4) [17]. La reconnaissance des sons par l'ordinateur nécessite, de la même manière que l'oreille humaine, l'extraction d'un ensemble de descripteurs audio. Ces descripteurs sont équivalents aux indices acoustiques identifiés par l'oreille humaine [18]. La reconnaissance est accomplie en utilisant des modèles de classification statistique. La localisation de l'endroit exacte du son est achevée en utilisant les techniques de localisation des sources sonores [19].

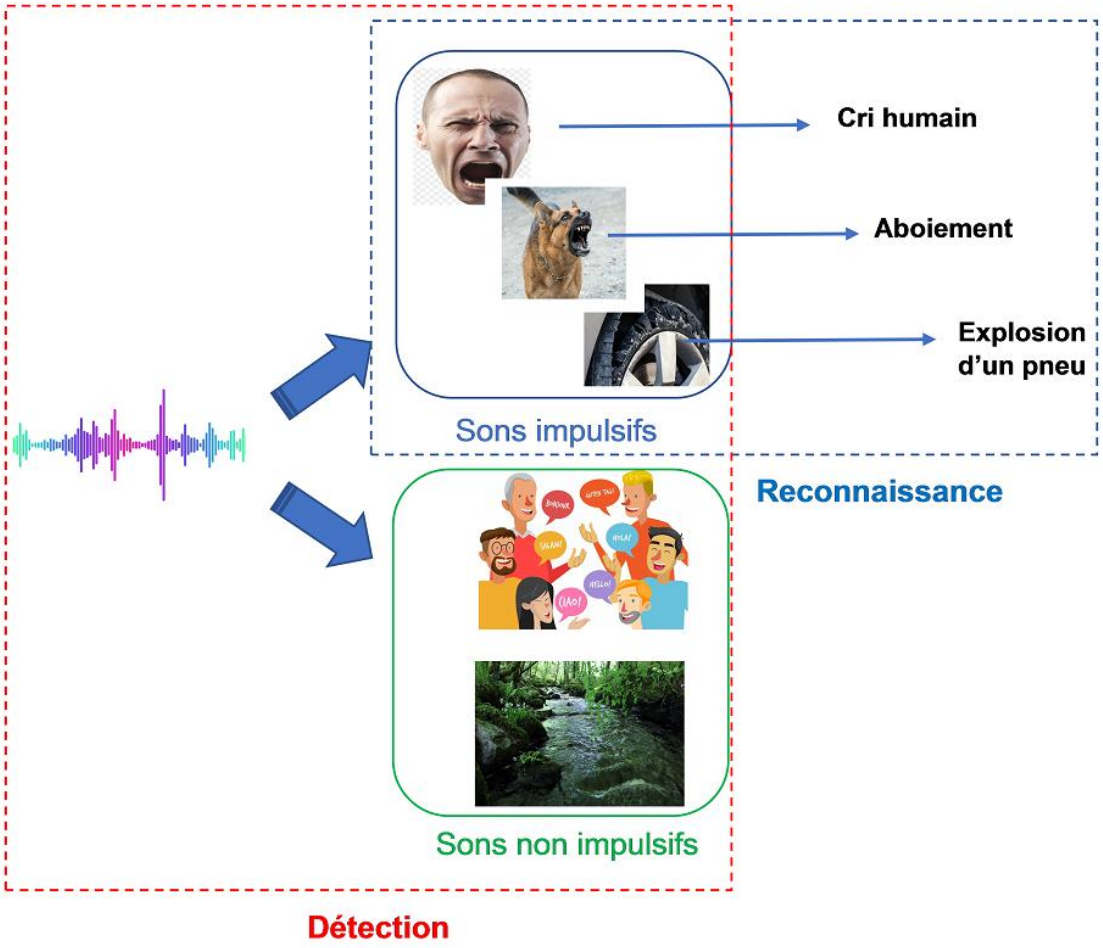

Figure 2.4: Exemples de détection et de reconnaissance des sons impulsifs

2.3.3 Etat de l'art sur les systèmes de reconnaissance des événements audio

Nous présentons dans cette section quelques travaux scientifiques sur la reconnaissance des évènements audio.

S. Souli et Z. Lachiri [20] ont proposé une nouvelle approche de reconnaissance des sons environnementaux pour les applications de surveillance médicale. Le système proposé a pour objectif de reconnaitre les sons environnementaux spécifiques (sonnettes de porte, signaux d'alarmes, etc.) en vue d'informer les personnes âgées ayant des troubles auditives. L'approche proposée est basée sur les attributs scattering et l'analyse en composantes principales. Les *support Vector Machines* (SVMs) ont été utilisées pour la classification des sons. D'après les résultats obtenus, il a été reporté que l'approche de classification par les SVMs est bien adaptée aux tâches de reconnaissance des sons environnementaux.

A. Rabaoui et al. [21] ont proposé une méthode de classification des sons à des fins de télésurveillance. Ils se sont intéressés aux sons suivants : cris de secours, coups de fusils, bris de glace, explosions, claquements de portes, aboiement de chiens, sonneries de téléphones, voix d'enfants et sons des machines. L'approche développée utilise des attributs basés sur une analyse en ondelettes et une procédure de classification à base de plusieurs SVMs à une seule classe. Les résultats obtenus montrent que cette approche a atteint de bonnes performances de reconnaissance.

La détection des changements dans les flux de données a été abordée par S. Rovetta et al. [22]. Ils se sont intéressés à la surveillance du trafic audio en vue de détecter : (a) les non-événements (bruit de fond), les événements non dangereux et dangereux, et (b) les non-accidents et accidents. Pour atteindre cet objectif, les chercheurs proposent un classificateur d'ensemble basé sur les SVMs à une classe afin de séparer d'abord les valeurs aberrantes (outliers) des données normales. Par la suite, ils proposent un réseau de neurones profond pour classer les données liées aux événements. Enfin, les résultats de la détection des valeurs aberrantes et des sorties de classification sont agrégés de telle sorte que les valeurs aberrantes sont considérées comme une nouvelle classe, a priori inconnue par le classificateur profond. Les chercheurs ont obtenu de bonnes performances de classification en utilisant des données réelles.

2.4 Conclusion

Nous avons mis en évidence dans ce chapitre les différentes générations des systèmes de surveillance. Nous avons montré l'apport de la modalité audio pour la surveillance. Nous présentons dans le troisième chapitre la méthode que nous proposons pour reconnaitre les événements audio dans un parking.

Chapitre 3 : Reconnaissance des événements audio

3.1 Introduction

Dans ce chapitre, nous présentons deux sections importantes : (i) la méthode de reconnaissance des évènements sonores, et (ii) les métriques utilisées pour évaluer cette méthode. Nous nous focalisons dans la première section sur le corpus sonore, l'extraction d'attributs et leurs normalisations ainsi que le schéma de classification à base de la programmation dynamique. Notre objectif est de comparer les performances de reconnaitre des événements audio en utilisant deux techniques d'extractions d'attributs : (i) les MFCCs et (ii) les LPCs. Dans la deuxième section, nous présentons : (i) la matrice de confusion, (ii) le taux de bonne classification, (iii) la précision et (iv) le rappel. Nous terminons ce chapitre par une conclusion.

3.2 Méthode proposée pour la reconnaissance des évènements audio

La reconnaissance des événements audio (En anglais : *Sound Events Recognition* (SER)) est le processus qui consiste à classer le signal audio en différentes catégories [23]. Cette technologie est utilisée dans plusieurs domaines d'applications tels que la surveillance de la circulation sur les routes, la sécurité publique et commerciale, et l'étude environnementale.

Dans notre projet, nous nous focalisons sur l'application de la SER pour la surveillance des parkings. Plus particulièrement, nous nous intéressons à la reconnaissance de trois évènements sonores, à savoir : (ii) cris humains, (ii) bris de glace, (iii) alarmes de voitures,

Le schéma de classification que nous proposons est basé sur la déformation temporelle dynamique (en anglais : *Dynamic Time Warping* (DTW)). Notre choix est motivé par le fait que nous construisons un modèle de reconnaissance en utilisons peu de données. Il est bien connu que lorsque le nombre de données est trop petit, les méthodes de type « template matching » sont plus appropriées que ceux à base de l'apprentissage (*Deep or schallow learning*) [24].

Nous utilisons deux techniques pour l'extraction d'attributs : (i) les MFCCs et les (ii) LPCs. Les MFCCs sont basés sur le processus de la perception auditive humaine tandis que les LPCs sont estimés à partir du modèle phonatoire de production de la parole (le modèle source filtre) [25].

Notre objectif est de comparer les performances de la reconnaissance en utilisant ces deux techniques. La Figure 3.1 schématise notre démarche.

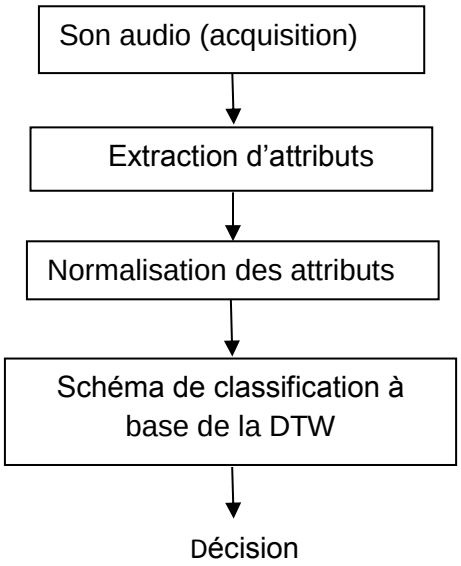

Figure 3.1: Méthode globale de reconnaissance des évènements sonores

3.2.1 Acquisition et prétraitement des sons

Le système de reconnaissance des évènements sonores est construit à partir des sons qui caractérisent les trois catégories citées ci-dessus. Nous avons téléchargé ces sons à partir de youtube [26] et sounddogs [27]. Nous avons par la suite construit une petite base de données (corpus) pour le tester et évaluer la méthode proposée. Une description détaillée sur les sons téléchargés ainsi que la procédure de leur prétraitement est fournie dans le chapitre 4.

3.2.2 Extraction des attributs

Le processus d'extraction d'attributs acoustiques consiste à calculer un vecteur de mesures qui représente l'évolution du signal au cours du temps [28]. [29]. Le choix d'attributs est d'une importance cruciale dans la reconnaissance des signaux [30]. De nombreuses techniques d'extraction de caractéristiques sont disponibles dans la littérature. Nous nous focalisons dans notre étude aux MFCCs et LPCs.

a) MFCCs

Les coefficients les plus utilisés pour extraire les caractéristiques spectrales des signaux de la parole sont les MFCCs [29]. L'échelle fréquentielle utilisée pour l'extraction des MFCCs est non-linéaire (échelle de Mel). Ces coefficients sont généralement appliqués en reconnaissance

automatique de la parole et du locuteur. Cependant, dans notre étude, nous utilisons ces coefficients pour extraire les informations fréquentielles des signaux de l'environnement.

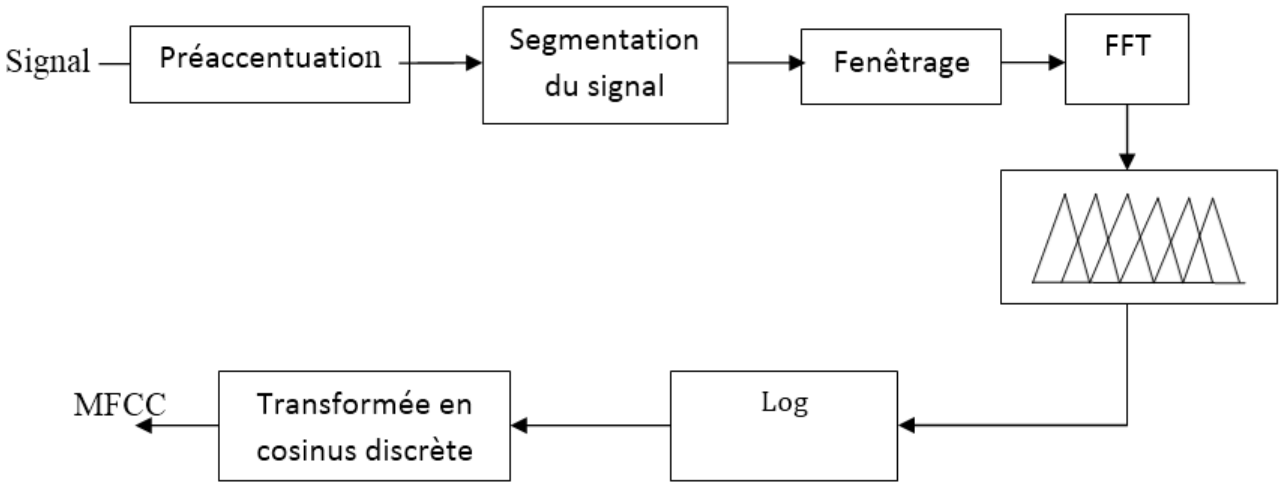

Figure 3.2: Calcul des MFCCs [31]

Les étapes nécessaires pour le calcul des coefficients MFCCs sont (Figure 3.2).

1) -Préaccentuation

Le signal s(n) est filtré par un filtre passe-haut :

$$s1\,(n) \;=\; s\,(n) \;-\; a \;*\; s\,(n-1) \tag{3.1}$$

s1(n) est le signal de sortie. La valeur de la constante « a » est comprise entre 0,9 et 1,0.

Le but de la préaccentuation est d'amplifier la partie haute fréquence du signal [31].

2) Segmentation du signal

Dans cette étape, le signal est divisé en petites trames d'échantillons avec un recouvrement de 50%. La longueur de la trame doit être comprise entre 10 et 40 ms pour garantir la quasi-stationnarité du signal [32].

3) Fenêtrage

Chaque trame doit être multipliée par une fenêtre de pondération afin de conserver la continuité du premier et du dernier point de la trame [31]. La fenêtre la plus utilisée est celle de Hamming. Elle est définie par l'équation suivante :

$$w\,(n, a) = \;(1 - a) - \; a\cos\left(\frac{2\pi n}{(N-1)}\right), \;\; 0 \;\leqq\; n \;\leqq\; N-1 \tag{3.2}$$

Si le segment d'un signal donné est noté s1 (n), n = 0,… N-1 (N est le nombre des échantillons) ; alors le segment après fenêtrage est calculé par la formule suivante :

$$s2(n)=s1\ (n) * w\ (n,a) \tag{3.3}$$

4) – Transformée de Fourier discrète

Le but de la transformée de Fourier discrète (calculée par l'algorithme FFT (En anglais : *Fast Fourier Transform* (FFT)) est de convertir le signal s2(n) du domaine temporel au domaine fréquentiel [32] :

$$Y(f) = FFT(s2(n)) \tag{3.4}$$

Où:

Y (f): est la transformée de Fourier discrète de s2(n),

5)-Filtres passe-bande triangulaires

Un ensemble de filtres passe-bandes triangulaires est utilisé pour approximer la résolution des fréquences perçues par l'oreille humaine [32]. Cette résolution est appelée échelle de Mel.

L'évolution de cette résolution est linéaire jusqu'à 1000 Hz et logarithmique par la suite. Pour passer de l'échelle de fréquence linéaire à l'échelle de Mel, nous utilisons l'équation suivante [32] :

$$f_{mel} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{3.5}$$

Où f est la fréquence en Hz.

Le spectre du signal à la sortie est le résultat du filtrage triangulaire.

6) - Logarithme

La prochaine étape consiste donc à calculer le logarithme de l'amplitude du spectre [33].

7) –Transformée en cosinus discrète

Dans cette étape, le logarithme du spectre est transformé au domaine Cepstral en utilisant la DCT (En Anglais : *Discrete Cosine Transform*). Le résultat de la conversion donne les coefficients MFCCs [31].

Les dérivées des MFCCs peuvent aussi être calculées.

- Les dérivées premières montrent la vitesse de variation de ces vecteurs dans les temps,
- Les dérivées deuxièmes donnent des informations sur l'accélération du signal.

b) LPCs

Le calcul des coefficients LPCs est basé sur la théorie du codage prédictive de la parole [30]. Dans notre étude, nous utilisons ces coefficients pour extraire les informations fréquentielles des signaux de l'environnement. L'hypothèse principale est que le signal audio peut être modélisé par un processus linéaire [34]. Le schéma de principe de LPC est illustré sur la Figure 3.3.

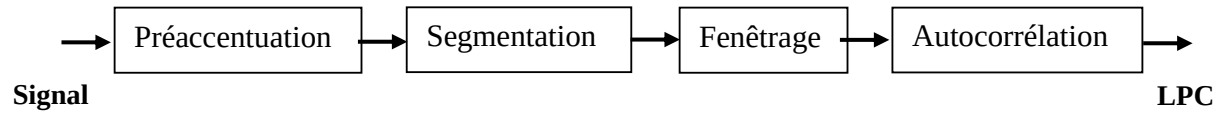

Figure 3.3: Calcul des LPCs

Le calcul des LPCs est effectué comme suit :

1) -Préaccentuation, segmentation et Fenêtrage : Il s'agit des mêmes opérations décrites dans la section précédente pour le calcul des MFCCs.

2)- Calcul des LPCs : Deux méthodes peuvent être utilisées pour estimer ces coefficients :

- Auto-corrélation,
- Covariance,

Nous nous intéressons dans notre travail à la méthode d'auto-corrélation. Cette méthode considère un échantillon du signal s(n) au temps n, et l'approximera par une combinaison linéaire des p précédents échantillons de la manière suivante [35] :

$$s(n) = a_1 s(n-1) + a_2 s(n-2) + \ldots\ldots + a_p s(n-p) \quad (3.6)$$

Où a_1,..... a_p sont des coefficients constants.

L'équation ci-dessus peut être transformée, en incluant un terme d'excitation G×u(n):

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + G \times u(n) \quad (3.7)$$

Où G est le gain et u(n) est l'excitation normalisée.

En calculant la transformée en Z de l'équation (3.7), nous obtenons :

$$S(z) = \sum_{i=1}^{p} a_i z^{-i} S(z) + G \times U(z) \quad (3.8)$$

Et par conséquent, la fonction de transfert H(z) devient :

$$H(z) = \frac{S(z)}{G \times U(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_1 z^{-1}} = \frac{1}{A(z)} \quad (3.9)$$

Cela correspond à la fonction de transfert d'un filtre numérique variable dans le temps.

Un prédicteur linéaire avec des coefficients a_k est défini comme suit (ref) :

$$\mathrm{P(z)} = \sum_{\mathrm{k=1}}^{\mathrm{p}} \mathrm{a_k z^{-k}} \tag{3.10}$$

Dont la sortie est :

$$\tilde{\mathrm{s}}(\mathrm{n}) = \sum_{\mathrm{k=1}}^{\mathrm{p}} \mathrm{a_k s(n-k)} \tag{3.11}$$

L'erreur de prédiction e(n) est définie comme :

$$e(\mathrm{n}) = \mathrm{s(n)} - \tilde{\mathrm{s}}(\mathrm{n}) = \mathrm{s(n)} - \sum_{\mathrm{k=1}}^{\mathrm{p}} \mathrm{a_k s(n-k)} \tag{3.12}$$

Le but principal est maintenant d'obtenir l'ensemble de coefficients a_k qui minimise l'erreur quadratique de prédiction dans un court segment du signal (généralement une trame de 10 à 20 ms). L'erreur de prédiction moyenne à court terme par trame est définie comme suit :

$$\mathrm{E_n} = \sum_{\mathrm{m}} \mathrm{e_n^2\,(m)} = \left[\mathrm{s_n(m)} - \sum_{\mathrm{k=1}}^{\mathrm{p}} \mathrm{a_k s_n(m-k)}\right]^2 \tag{3.13}$$

Où $s_n(m)$ est un segment du signal sélectionné dans le voisinage de l'échantillon n :

$$\mathrm{s_n(m)} = \mathrm{s(m+n)} \tag{3.14}$$

La minimisation de l'équation (3.13) nous conduit directement vers la forme matricielle suivante:

$$\begin{bmatrix} R_n(0) & R_n(1) & \cdots & R_n(2) & R_n(p-1) \\ R_n(1) & R_n(0) & \cdots & R_n(1) & R_n(p-2) \\ R_n(2) & R_n(1) & \cdots & R_n(0) & R_n(p-3) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ R_n(p-1) & R_n(p-2) & \cdots & R_n(p-3) & R_n(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \cdots \\ a_p \end{bmatrix} = \begin{bmatrix} R_n(1) \\ R_n(2) \\ R_n(3) \\ \cdots \\ R_n(p) \end{bmatrix} \tag{3.15}$$

Pour plus de détails sur le passage entre l'équation (3.14) et l'équation (3.15), le lecteur est invité à lire la référence [35]. L'équation 3.15 peut être résolue en utilisant plusieurs méthodes. Parmi ces méthodes on peut citer : (i) l'inversion matricielle et (ii) la méthode de Durbin [35].

3.2.3 Normalisation d'attributs

La normalisation est l'opération de mise à l'échelle des valeurs numériques d'attributs. Il existe plusieurs méthodes de normalisation dans la littérature. On peut citer : (i) la normalisation min-max, et (ii) le zscore [28].

Dans notre travail, on a utilisé la deuxième méthode qui consiste à :

- Calculer la moyenne et l'écart type de chaque attribut,
- Soustraire la moyenne de chaque attribut et le normaliser par rapport à son écart type,

$$x' = \frac{x - \overline{x}}{\sigma} \tag{3.16}$$

x est l'attribut sans normalisation, $\overline{x}$ est sa moyenne, et σ est son écart type,

x'est l'attribut normalisé.

3.2.4 Schéma de classification à base de la déformation temporelle dynamique

La méthode de classification que nous adoptons est basée sur la DTW vu que nous cherchons à classer les séries temporelles (signaux audio).

Cette section est structurée en deux parties. La première partie donne un aperçu sur la classification des séries temporelles. Une attention particulière est accordée au calcul des distances. La deuxième partie présente la méthode proposée pour la classification des séquences.

a) Classification des séries temporelles

Pour un ensemble donné de séries temporelles divisées en plusieurs classes avec étiquettes, le problème de classification consiste à assigner la série temporelle de test à la classe correspondante [36] [37] [38].

a.1) Calcul de distances

Nous introduisons dans ce qui suit les distances utilisées pour la classification des séries temporelle. Nous nous focalisons principalement sur la distance Euclidienne et la déformation temporelle dynamique.

- **<u>Distance euclidienne</u>**

Supposons qu'il existe trois séries temporelles ordonnées et discrètes A = {a_1, a_2, a_3, ..., a_n}, B = {b_1, b_2, b_3, ..., b_n} et C = {c_1, c_2 , c_3, ..., c_n}, où chaque valeur des points temporels a_i, b_j, $c_k \in R$, et $1 \leq i, j, k \leq n$.

La distance euclidienne entre A et B est donnée comme suit :

$$\mathrm{ED}\,(\mathrm{A},\mathrm{B}) = \sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{ai} - \mathrm{bj})^2} \tag{3.17}$$

Deux séries temporelles sont similaires si leur distance est plus courte.

Par exemple :

Considérons les deux séries temporelles A et B tel que :

$A = \{1,2,3,2,2\}$, et $B = \{2,2,3,4,4\}$;

Soit C= {1,1,2,2,3}, une série temporelle de test. Si on cherche à trouver la série temporelle la plus similaire à C, on calcul les distances Euclidiennes ED (C, A) et ED (C, B).

Les résultats trouvés sont :

$ED\ (C, A) = \sqrt{3}$

$ED(C, B) = \sqrt{8}$

Vu que ED (C, A) est inférieur à $ED(C, B)$, alors, C est plus proche, ou plus similaire à A.

La distance Euclidienne est une mesure directe de calcul de la distance entre les séries temporelles. Cependant, cette mesure n'est pas utilisée en pratique vu que la longueur de deux séries temporelles données peut ne pas être la même. Dans ce cas, si ces deux séries sont équivalentes et que l'une est légèrement modifiée le long de l'axe des temps, alors, ces deux séries peuvent être jugées différentes. De plus, la distance Euclidienne n'est pas invariante aux données distordues qui sont très répandues dans de nombreux domaines [36] [37].

- **DTW**

La DTW est très utilisée pour la classification des séries temporelles. Néanmoins, sa complexité algorithmique est très élevée par rapport à la distance Euclidienne.

Soit $A = \{a_1, a_2, ..., a_m\}$ et $B = \{b_1, b_2, ..., b_n\}$ deux séries temporelles. La variable $DTW(i, j)$ désigne la distance entre $A_{1...i}$ et $B_{1...j}$.

La formule de programmation dynamique pour le calcul de la DTW est donnée par l'équation suivante :

$$DTW(i,j) = \begin{cases} 0 & si\ i = 0\ et\ j = 0, \\ \infty & si\ i = 0\ où\ j = 0, et\ i \neq j \\ dis(ai, bj) + \min \begin{cases} DTW(i-1, j) \\ DTW(i, j-1) \\ DTW(i-1, j-1) \end{cases} & si\ 1 \leq i \leq m\ et\ 1 \leq j \leq n. \end{cases} \quad (3.18)$$

La variable $dis(a_i, b_j)$ représente la distance entre a_i et b_j.

B / A		3	7	4	1	3	2	1	7
	0	∞	∞	∞	∞	∞	∞	∞	∞
2	∞	1	6	8	9	10	10	11	16
9	∞	7	3	8	16	15	17	18	13
8	∞	12	4	7	14	19	21	24	14
8	∞	17	5	8	14	19	25	28	15
5	∞	19	7	6	10	12	15	19	17
4	∞	20	10	6	9	10	12	15	18
2	∞	21	15	8	7	8	8	9	14
1	∞	23	21	11	7	9	9	8	14
5	∞	25	23	12	11	9	12	12	10

Figure 3.4: Calcul de la DTW (les cellules grises représentent le chemin de l'alignement temporel dynamique) [38].

La Figure 3.4 montre un exemple de calcul de la DTW entre les deux séries suivantes : A= {2,9,8,8,5,4,2,1,5}, et B= {3,7,4,1,3,2,1,7}.

Les longueurs des deux séries ne sont pas les mêmes. La distance DTW entre A et B est M [9,8] = 10 (la solution optimale).

Le *k*ième élément du chemin de déformation W est défini par W*k* = (i, j). Le chemin global de déformation est :

$W = \{w1, w2, \dots, wk, \dots, wk'\}$, Où max $(m, n) \leq k \leq m + n - 2$.

Il y a trois contraintes pour trouver le chemin de déformation.

- Conditions aux limites : $w1 = (1,1)$et $wk' = (m, n)$. En d'autres termes, le chemin de déformation commence et se termine dans les cellules opposées de la diagonale de la matrice.
- Continuité **:** Etant donné $w_k = (a, b), w_{k-1} = (a', b')$ et $w_k \neq w_{k-1}$, où $a - a' \leq 1$ et $b - b' \leq 1$: Il est indiqué que deux éléments adjacents quelconques dans le chemin de déformation sont des cellules adjacentes (y compris des cellules adjacentes en diagonale).
- Monotonie **:** Etant donné $w_x = (a, b)$, $w_y = (a', b')$ où $1 \leq x \leq y \leq k'$, alors $a' - a \geq 0$ et $b' - b \geq 0$. Cela signifie que le chemin de déformation ne retourne pas.

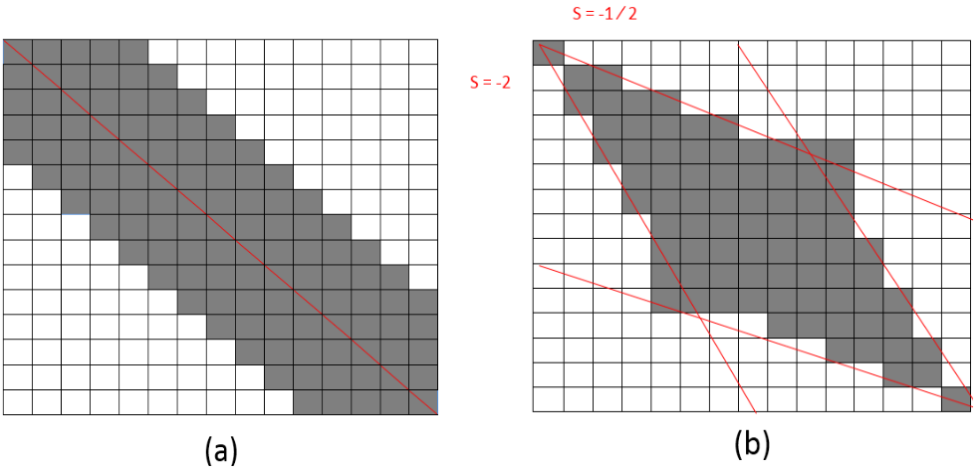

Figure 3.5: Fenêtres de déformation de la DTW. (a) Bande de Sakoe-Chiba, délimitée par | i-j | <r = 5. (b) Parallélogramme d'Itakura, délimité par des pentes S = -2 et S = -1/2.

- **Autres variantes de la DTW**

Le calcul de la distance à base de la DTW prend du temps lorsque la taille des données est importante [36] [37]. Pour faire face à ce problème, la notion de déformation temporelle dynamique avec fenêtre a été proposée par Itakura [36] en 1975, et Sakoe et Chiba [37] en 1978. L'algorithme d'Itakura [36] est basé sur la fonction de pente S, et le chemin de déformation est délimité par deux pentes S et 1 / S. Sakoe et Chiba [37] ont utilisé une diagonale de largeur fixe.

La principale différence entre ces méthodes et la DTW classique est que ces méthodes utilisent des contraintes de déformation supplémentaires qui peuvent être réglées de 100% à 0%.

Le but des contraintes est de permettre au chemin de déformation d'être plus proche de la diagonale et d'éviter le chemin indésirable. Supposons que la taille des fenêtres de déformation est r. Alors, le chemin n'est autorisé que dans la largeur r, c'est-à-dire. | i-j | <r.

Les contraintes globales bien connues, proposées par Sakoe et Itakura sont illustrés sur la Figure 3.5. Dans notre travail, nous avons utilisé la méthode de Sakoe-Chiba.

b) Schéma de classification proposé

Nous nous intéressons dans ce projet à la reconnaissance des événements sonores. L'étape de détection des sons impulsifs a été déjà étudiée et implémentée dans [15].

Le schéma de classification que nous proposons consiste à utiliser un codebook de référence pour assigner le son de test à la classe la plus probable. Le début exact du son est considéré connu. Le codebook est constitué de 30 triplets différents.

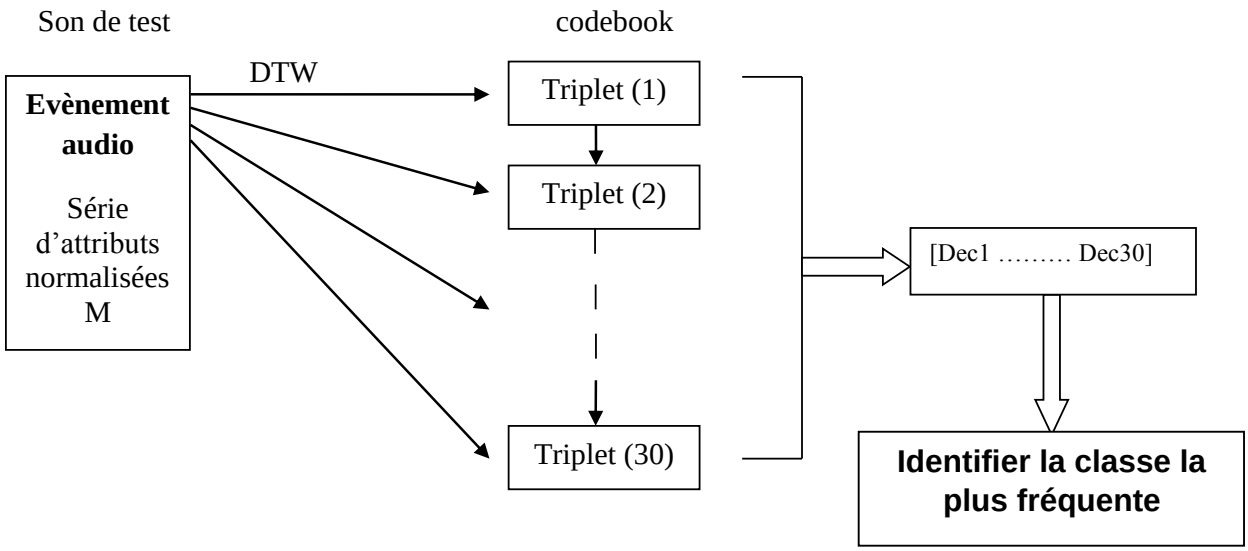

Figure 3.6: Schéma de classification.

Chaque triplet (T(i)) est composé de trois matrices correspondant aux valeurs numériques d'attributs extraits à partir des trois séquences de référence, à savoir : (i) alarmes de voitures, (ii) bris de glace et (iii) cris des humains. Le calcul des distances est effectué par la DTW. La procédure de test est donnée comme suit (Figure 3.6) :

(i) lire le son de test (évènement audio),

(ii) introduire la durée du segment à traiter « D »,

(iii) introduire le nombre d'attributs acoustiques « NF » (MFCCs ou LPCs),

(iv) calculer et normaliser les valeurs numériques d'attributs acoustiques (MFCCs ou LPCs),

(v) constituer la matrice d'attributs M à partir de ces attributs,

- le nombre de lignes correspond à la longueur de la séquence,
- le nombre de colonnes correspond au nombre d'attributs,

3.3 Métriques d'évaluation

Cette section présente un aperçu sur les métriques d'évaluations des méthodes de classification. Nous nous focalisions sur : (i) la matrice de confusion, (ii) la précision et le rappel, (iii) le taux de bonne classification et le taux d'erreur.

3.3.1 Matrice de confusion

La matrice de confusion pour un problème de classification à trois classes (A, B et C) est schématisée sur la Figure 3.7.

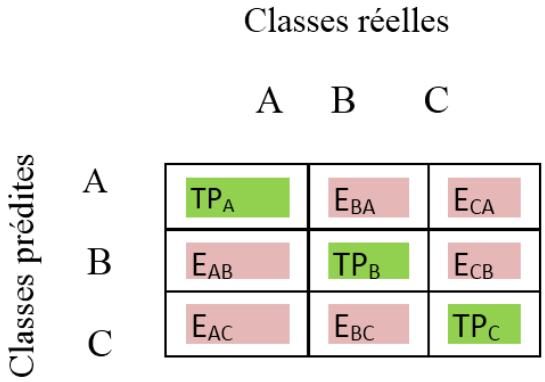

Figure 3.7: Structure de la matrice de confusion [40]
(La diagonale verte représente les prévisions correctes et la diagonale rose indique les prévisions incorrectes)

TP_A est le nombre d'échantillons qui sont véritablement positifs dans la classe A (en anglais, *True positive* (TP)), c'est-à-dire le nombre d'échantillons qui sont correctement classés dans la classe A, E_{AB} est le nombre d'échantillons de la classe A qui ont été incorrectement classés dans la classe B, c'est-à-dire les échantillons mal classés.

Ainsi, le faux négatif dans la classe A (en anglais, False Negative noté : $FN_{A,}$) est la somme de E_{AB} et E_{AC} ($FN_A=E_{AB}+E_{AC}$). Il correspond à la somme de tous les échantillons de classe A qui ont été incorrectement classés en classe B ou C. Simplement, le FN de toute classe située dans une colonne peut être calculé en ajoutant les erreurs dans cette classe / colonne.

Alors que pour le faux positif (en anglais : False Positive (FP)), toute classe prédite, qui se trouve dans une ligne, représente la somme de toutes les erreurs dans cette même ligne. Par exemple, le faux positif de la classe A est calculé comme suit, $FP_A=E_{BA}+E_{CA}$.

Plusieurs mesures peuvent être dérivées à partir de cette matrice. Généralement, le taux de bonne classification, la précision et le rappel sont utilisés pour des problèmes de classification où la répartition des échantillons est équilibrée (le même nombre d'échantillons pour chaque classe).

Cependant, d'autres métriques, tels que : F1 score, Macro F1, Macro-précision, Macro-rappel sont utilisées pour des cas déséquilibrés [41]. Nous nous focalisons dans notre projet à la première catégorie de mesures.

3.3.2 Rappel et précision

Le rappel d'un classificateur, en anglais « recall », représente le rapport des échantillons positifs correctement classés sur le nombre total d'échantillons positifs « P ». Cette métrique est aussi nommée « taux de vrais positifs » (en anglais, True positive rate (TPR)). On peut l'estimer en utilisant l'équation suivante :

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} = \frac{\text{TP}}{\text{P}} \tag{3.19}$$

La précision, en anglais « precision », est le rapport entre les prédictions positives correctes et le total des positifs prédits [40] [41]. Cette mesure est aussi appelée valeur prédictive positive. On peut l'estimer en utilisant l'équation suivante :

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{3.20}$$

Pour une classe donnée, les différentes combinaisons de rappel et de précision ont les significations suivantes [41]:

- Le Rappel est « élevé » + la précision est « haute » : la classe est parfaitement gérée par le modèle.
- Le Rappel est « faible » + la précision est « haute » : le modèle ne détecte pas correctement la classe, mais il est hautement fiable quand il le fait.
- Le Rappel est « élevé » + la précision est « faible » : la classe est bien détectée mais le modèle comprend également des points d'autres classes.
- Le Rappel est « faible » + la précision est « faible » : la classe est mal gérée par le modèle.

3.3.3. Taux de bonne classification

Le taux de bonne classification (en anglais Accuracy (Acc) est l'une des mesures les plus couramment utilisées pour évaluer les performances des classifieurs [40].

Cette métrique est définie comme le rapport entre les échantillons correctement classés et le nombre total d'échantillons :

$$\mathrm{Acc} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} = \frac{\mathrm{TP+TN}}{\mathrm{P+N}} \quad (3.21)$$

Où P et N indiquent respectivement le nombre d'échantillons positifs et négatifs, respectivement.

Le complément de cette la métrique est le taux d'erreur ou le taux de mauvaise classification (en anglais *Error Rate* (ER)). Cette métrique représente le nombre d'échantillons mal classés des classes. Elle est calculée comme suit :

$$ER=1-Acc= (FP+FN) / (TP+TN+FP+FN) \quad (3.22)$$

3.4 Conclusion

Dans ce chapitre, nous avons proposé une méthode de reconnaissance des sons de l'environnement pour le cas des bases de données (corpora) de petite taille « small size samples ». Par la suite, nous avons présenté les métriques nécessaires pour évaluer la méthode proposée. Dans le chapitre suivant, nous présenterons les résultats expérimentaux que nous avons obtenus.

Chapitre 4 : Résultats expérimentaux

4.1 Introduction

Dans ce chapitre, nous présentons trois sections importantes : (i) les logiciels de développement utilisés, (ii) l'implémentation de la méthode proposée et (iii) la comparaison entre les MFCCs et LPCs. Nous terminons ce chapitre par une conclusion.

4.2 Logiciels de développement

4.2.1 Environnement MATLAB

Ce projet a été réalisé à l'aide du logiciel MATLAB (Figure 4.1). C'est un langage performant pour le calcul technique. Il intègre le calcul, la visualisation et la programmation dans un environnement facile à utiliser où les solutions sont exprimées dans une notation mathématique familière. Nous avons utilisé la version 2015 (MATLAB R2015b) [42].

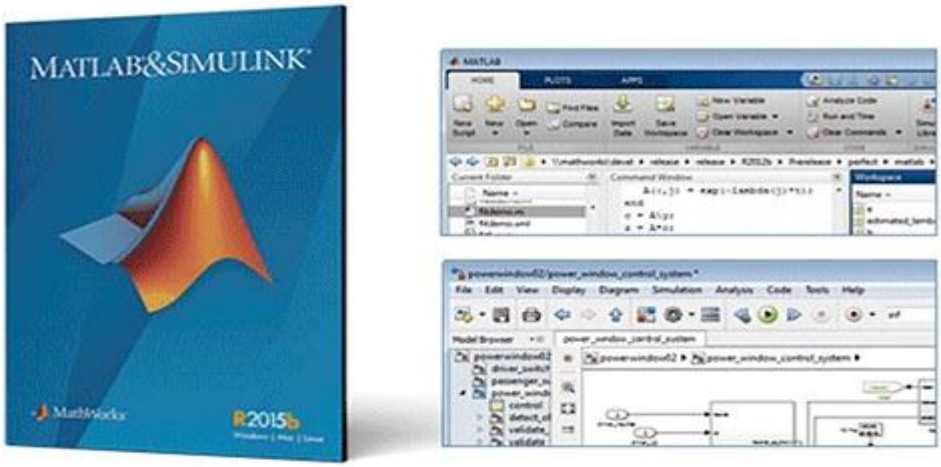

Figure 4.1 : Logiciel MATLAB

La toolbox nommée « voicebox » a été aussi utilisée pour extraire les attributs. C'est une boîte à outils de traitement de la parole composée de routines MATLAB écrites par Dr. Mike Brookes de l'Imperial College, Royaume-Uni [43]. Pour le calcul de la DTW, on a utilisé la fonction de Quan Wang, Kim de Rensselaer Polytechnic Institute [44].

4.2.2 GoldWave

GoldWave est un logiciel de montage audio numérique qui permet la lecture, le traitement et l'enregistrement des sons. Il a été développé par GoldWave Inc, lancé pour la première fois en Avril 1993 [45].

Il comprend un ensemble complet de fonctionnalités de traitement audio. Une fenêtre de contrôle indépendante fournit un accès direct aux périphériques audio. Il contient des commandes pour le rembobinage et l'avance rapide, l'enregistrement, le volume, la balance et la vitesse.

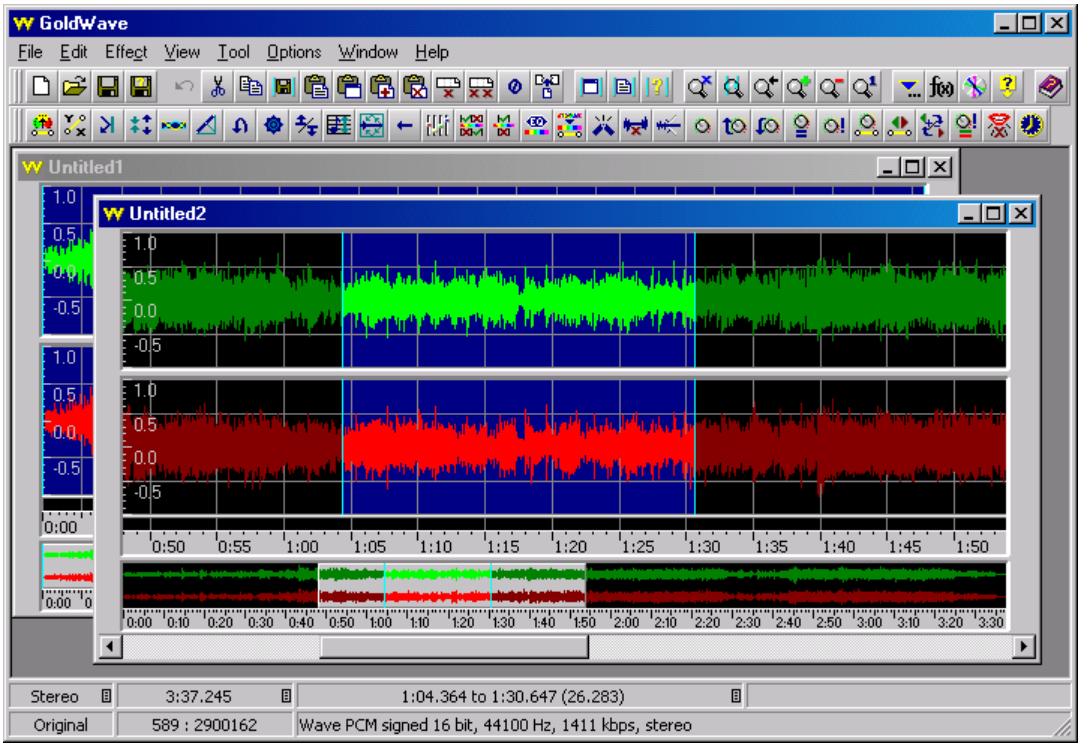

Figure 4.2 : GoldWave

4.3 Implémentation de la méthode proposée

4.3.1 Acquisition de données (Corpus de test)

La méthode que nous avons proposée pour la reconnaissance des évènements sonores a été testée sur un corpus de 150 sons. Comme nous l'avons indiqué dans le chapitre précédent, nous avons choisi trois catégories de sons : cris des humains, bris de glace, et alarme de voitures (Figure 4.3). Chaque classe contient 50 sons. Le corpus a été téléchargé gratuitement à partir des sites web : YouTube et Sounddogs [26] [27].

Alarme de voiture

Bris de glace

Cris humain

Figure 4.3 : Sons de la base de données

Le prétraitement du corpus a été effectué en utilisant le logiciel GoldWave. La fréquence d'échantillonnage des sons est de 44100 Hz. Nous avons éliminé manuellement les parties de silences. Nous avons ajusté le début et la fin de chaque son en cliquant sur le bouton : « *set Start marker* » et « set Finish marker » (Figure 4.4).

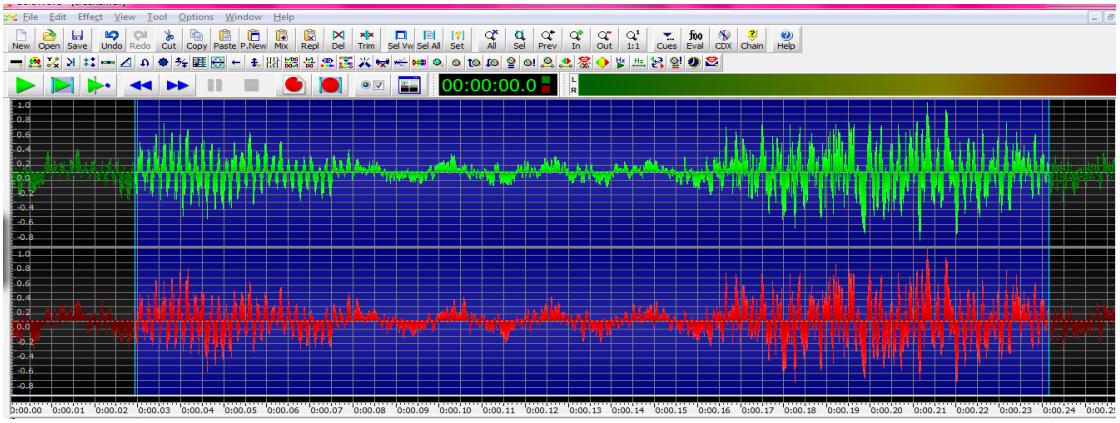

Figure 4.4 : Ajustement du début et de la fin d'un son

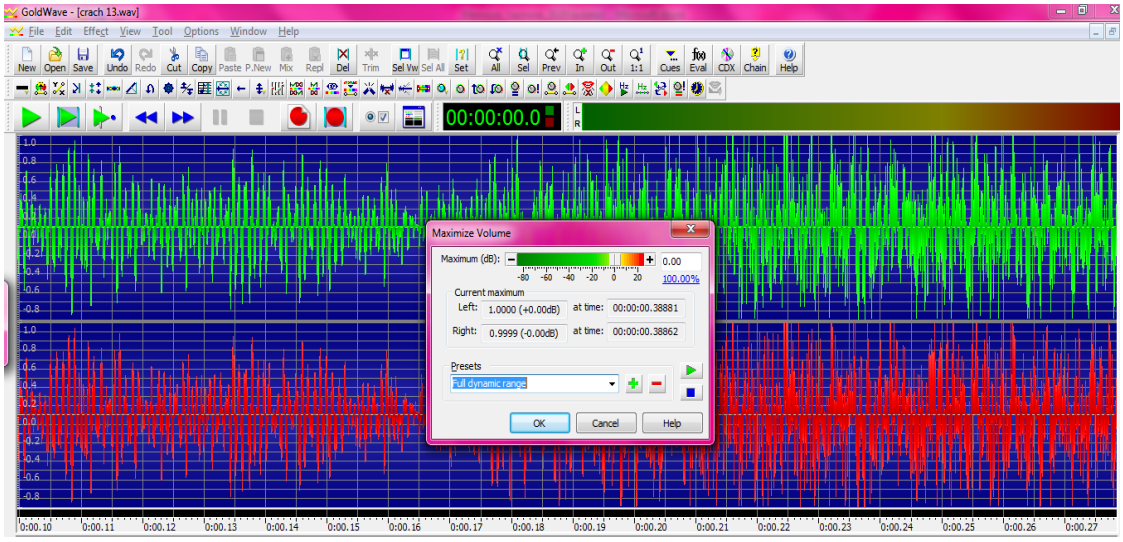

Figure 4.5 : Ajustement du volume d'un son

Pour obtenir des sons avec des amplitudes comparables, nous avons utilisé l'option « Dynamic Range Equalization » du Goldwave (Figure 4.5).

4.3.2 Procédure de test

Nous avons devisé aléatoirement notre corpus en deux groupes, un groupe pour le test et l'autre groupe pour construire les références (codebook). Parmi les 50 sons de chaque catégorie (cris des humains, bris de glace et alarmes de voitures), nous avons utilisé 30 sons pour construire le codebook et 20 sons pour effectuer le test. A cet effet, le codebook global, qui inclus les trois catégories, est composé de 90 sons tandis que le groupe global de test inclus 60 sons.

Pour concevoir une méthode de reconnaissance à complexité réduite, nous devons satisfaire la contrainte suivante :

- Trouver la durée minimale des sons de test qui aboutissent à un taux de reconnaissance élevé en utilisant le moins d'attributs possible,

Le schéma de reconnaissance que nous proposons sera intégrer par la suite dans une plateforme de surveillance audio existante au niveau du laboratoire ASM.

4.3.3 Réglage des durées

Nous présentons dans cette section les résultats obtenus pour le réglage des durées des sons. L'ajustement du nombre d'attributs est abordé dans la section suivante.

Nous avons varié les durées de tous les sons de notre corpus entre 0.1s et 0.6s (avec un pas de 0.1s). Pour chaque durée imposée, nous avons varié le nombre d'attributs de 2 à 22 (avec un pas de 1). Si nous imposons une durée fixe pour l'ensemble des sons du corpus, nous devons limiter ou augmenter la durée intrinsèque de chaque son.

A titre d'exemple, nous présentons par la suite la méthodologie de limitation et d'augmentation des durées des sons pour une valeur de 0.4s.

- **<u>Cas d'augmentation de la durée du signal</u>**

Considérons le cas d'un son avec une durée égale à 0.35s (Figure 4.6). Si nous voulons augmenter sa durée à 0.4s, nous devons rajouter à la fin du son un bruit blanc d'un écart type égal à $1\ 10^{-3}$ et d'une durée égale à 0.5s (Figure 4.7).

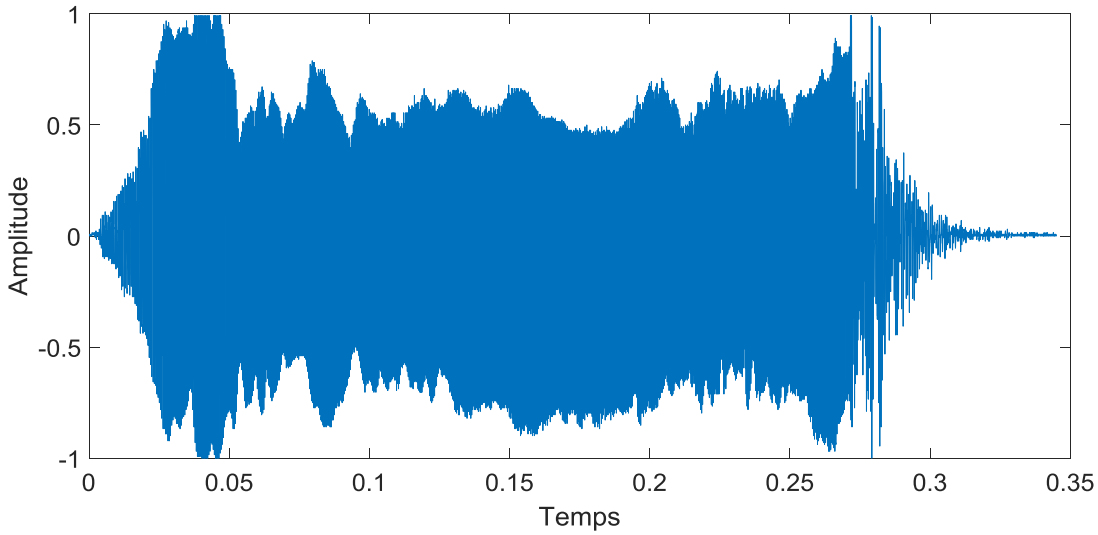

Figure 4.6 : Son d'une durée égale à 0.35s

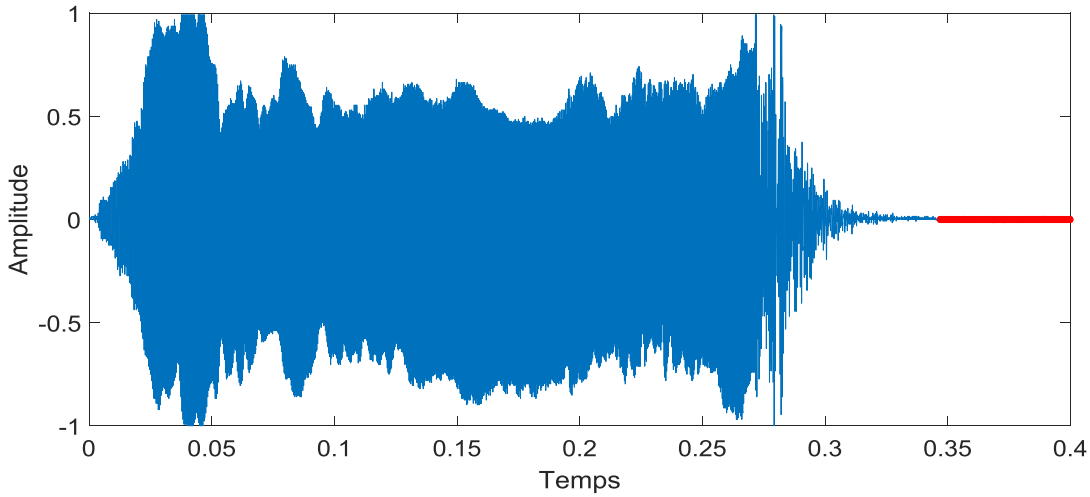

Figure 4.7 : Augmentation de la durée du son à une valeur égale à 0.4s

- **<u>Cas de limitation de la durée du son</u>**

Considérons le cas d'un son avec une durée égale à 0.6s (Figure 4.8). Si nous voulons limiter sa durée à 0.4s, nous devons éliminer quelques échantillons du signal (Figure 4.9).

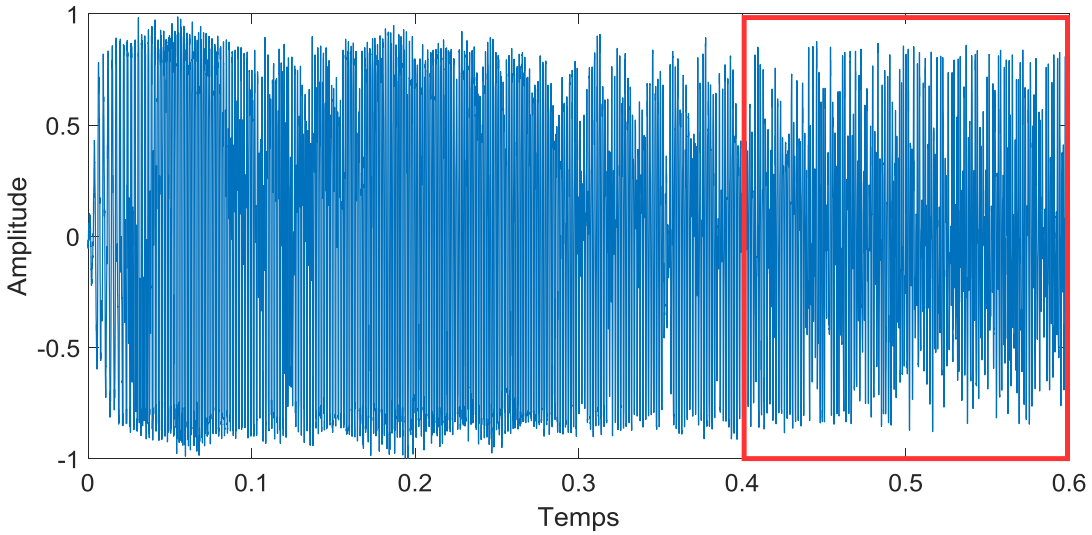

Figure 4.8 : Son d'une durée égale à 0.6s

Après la limitation de la durée, nous obtenons le signal suivant :

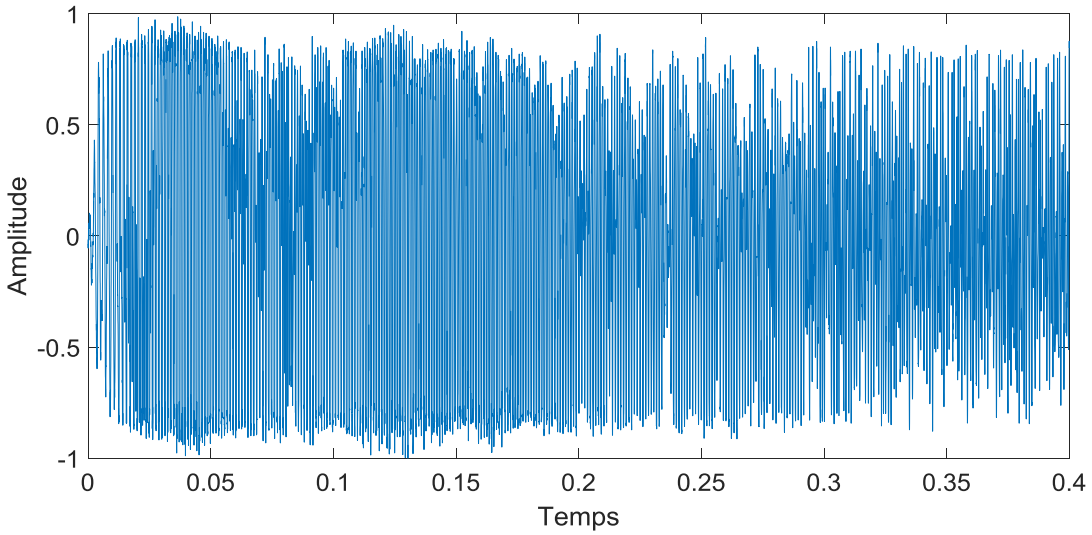

Figure 4.9 : Limitation de la durée d'un son à une valeur égale à 0.4s

4.3.4 Extraction d'attributs

Nous allons extraire deux catégories d'attributs : les MFCCs et les LPCs. Le calcul des MFCCs se fait avec et sans dérivées. Par contre, le calcul des LPCs se fait sans considération des dérivées.

- **Attributs MFCCs**

Pour une série temporelle (onde acoustique du son), le calcul des MFCCs est effectué en considérant les paramètres suivants :

- Le coefficient de préaccentuation est fixé à 0.95,
- La taille de la fenêtre d'analyse est de 30ms,
- Le recouvrement entre les fenêtres est de 15ms,
- Le nombre des bancs de filtre est de 32,

La Figure 4.10 montre un exemple de calcul de 12 MFCCs pour un cri humain.

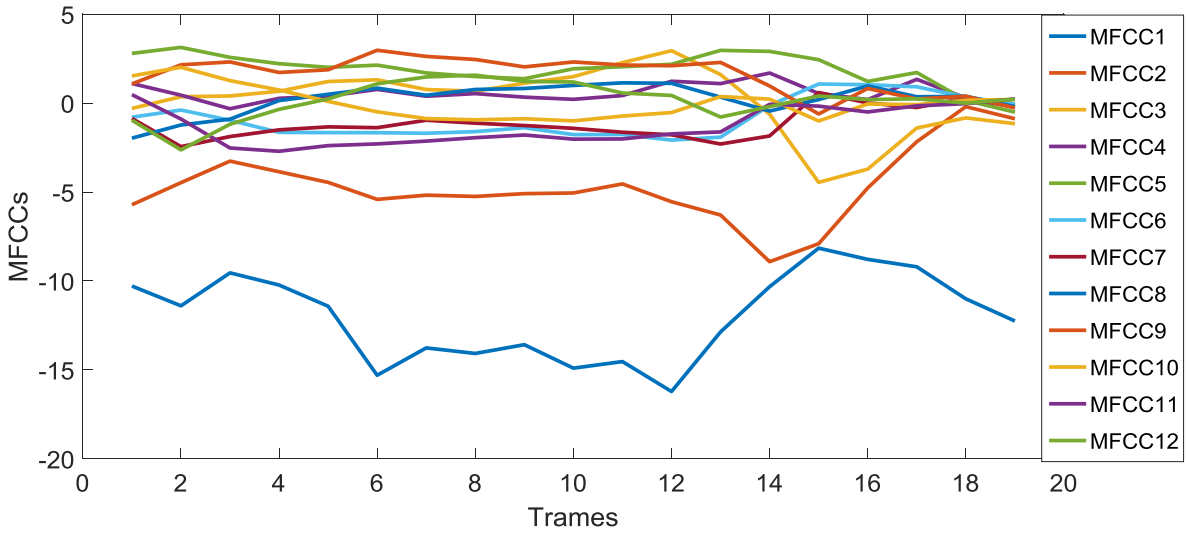

Figure 4.10 : Evolution des MFCCs d'un son en fonction des trames d'analyse (sans dérivées)

Dans une deuxième étape, nous calculons aussi les dérivées des MFCCs. La Figure 4.11 montre un exemple de calcul de 12 coefficients MFCCs en considérant les premières et deuxièmes dérivées. Le nombre total des coefficients dans ce cas est de 36 coefficients (12 MFCCs + 12 Delta MFCCs +12 Delta Delta MFCCs).

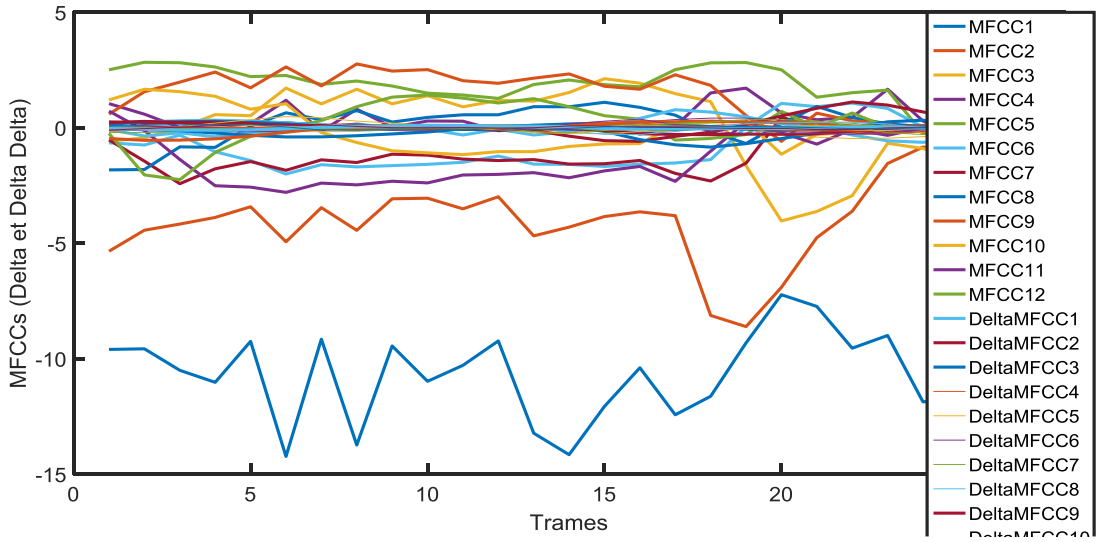

Figure 4.11 : Evolution des MFCCs d'un son en fonction des trames d'analyse (avec dérivées)

- **Attributs LPCs**

Pour une série temporelle (son), le calcul des LPCs est effectué en considérant les paramètres suivants :

- Le coefficient de préaccentuation est fixé à 0.95,
- La taille de la fenêtre d'analyse est de 30ms,
- Le recouvrement entre les fenêtres est de 15ms,

La Figure 4.12 montre un exemple de calcul de 13 LPCs pour un cri humain.

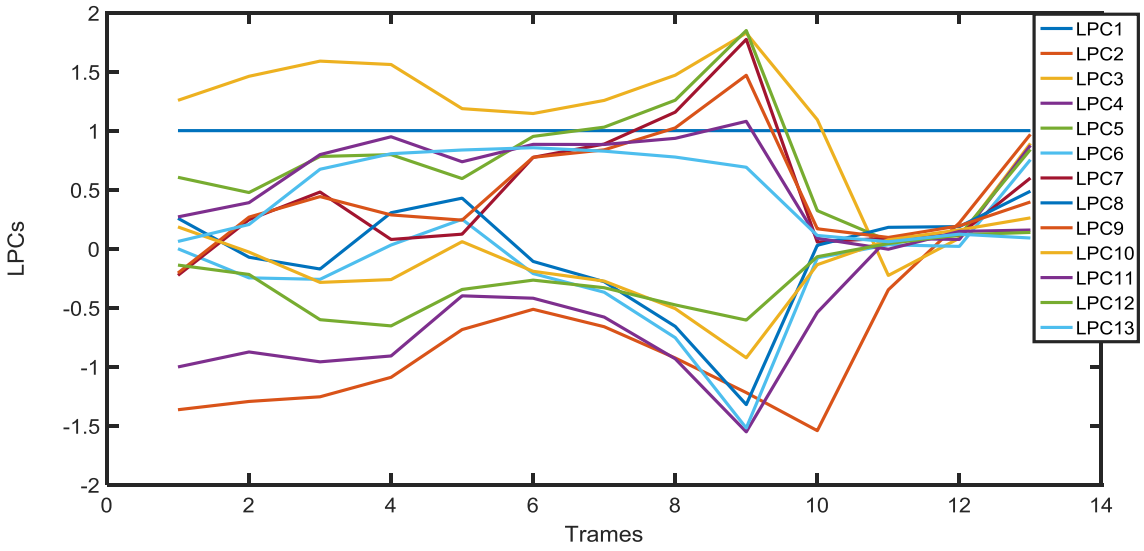

Figure 4.12 : Evolution des LPCs d'un son en fonction des trames d'analyse

La procédure de test consiste à varier le nombre d'attributs de 2 à 22 pour chaque durée imposée.

4.3.5 Normalisation d'attributs

L'objectif de cette étape est d'obtenir un ensemble d'attributs normalisés dont l'écart type et la moyenne sont respectivement un et zéros. Il est important de noter que cette opération est effectuée sur l'ensemble d'attributs du corpus (en utilisant tous les sons du corpus). Cependant, l'exemple suivant schématise l'opération sur un seul son (représenté par 12 MFCCs).

- Attributs avant la normalisation (Figure 4.13),
- Attributs après la normalisation (Figure 4.14).

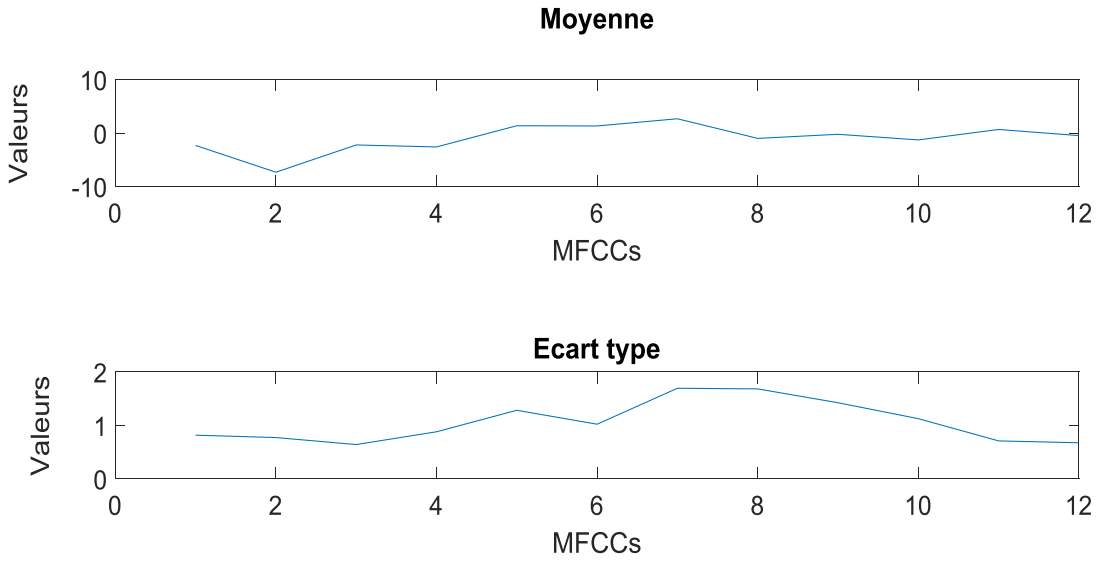

Figure 4.13 : Ecart type et moyenne des MFCCs avant l'opération de normalisation

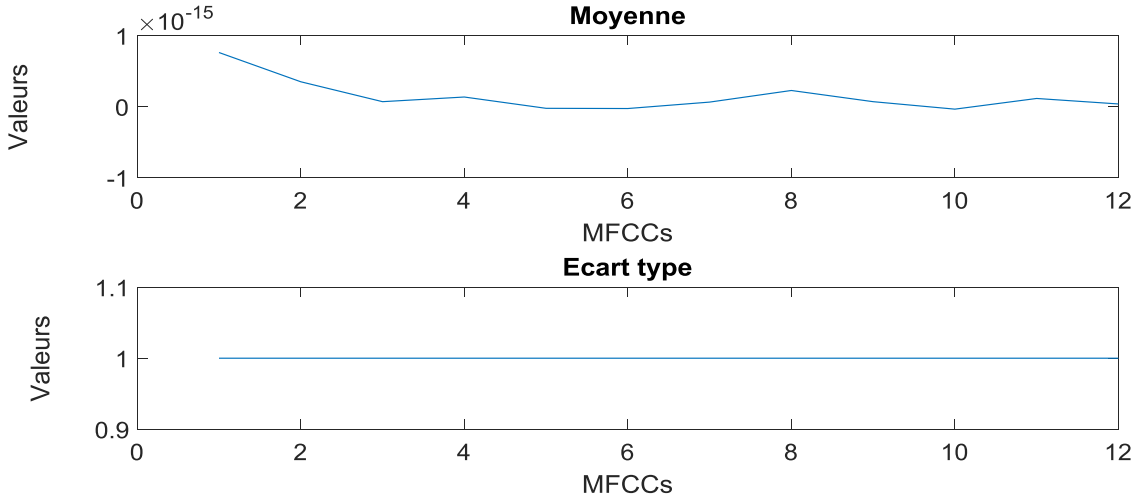

Figure 4.14 : Ecart type et moyenne des MFCCs normalisés

4.3.6 Classification/reconnaissance

Dans cette section nous allons décrire le schéma de classification mentionné dans le chapitre 3 (section 3.2.4). Nous présentons les résultats obtenus pour un seul son de test (cris humain).

- **Etape 1 : Lecture du son de test (Figure 4.15)**

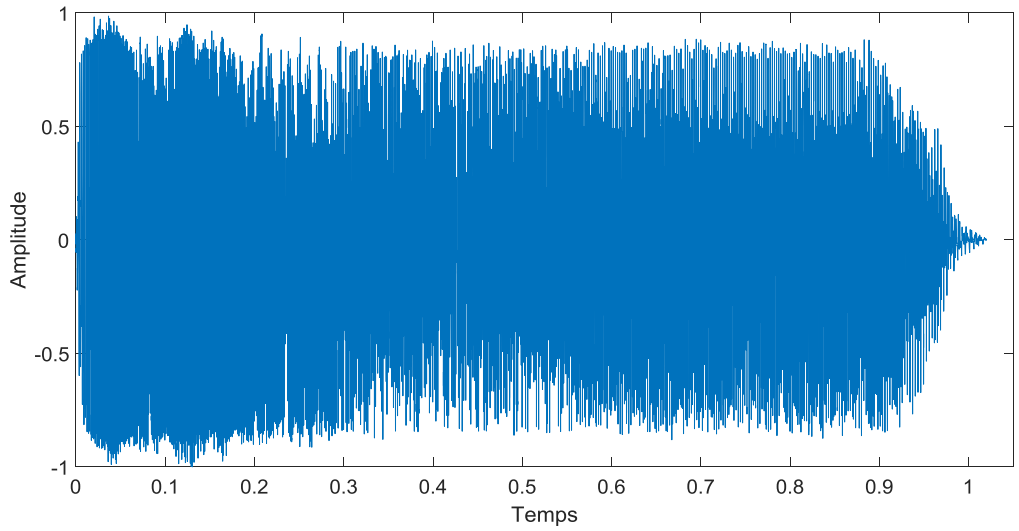

Figure 4.15 : Son de test original

- **Etape 2 : Introduction de la durée du segment à traiter**

Dans cet exemple, nous cherchons à limiter la durée du son à 0.2s (Figure 4.16).

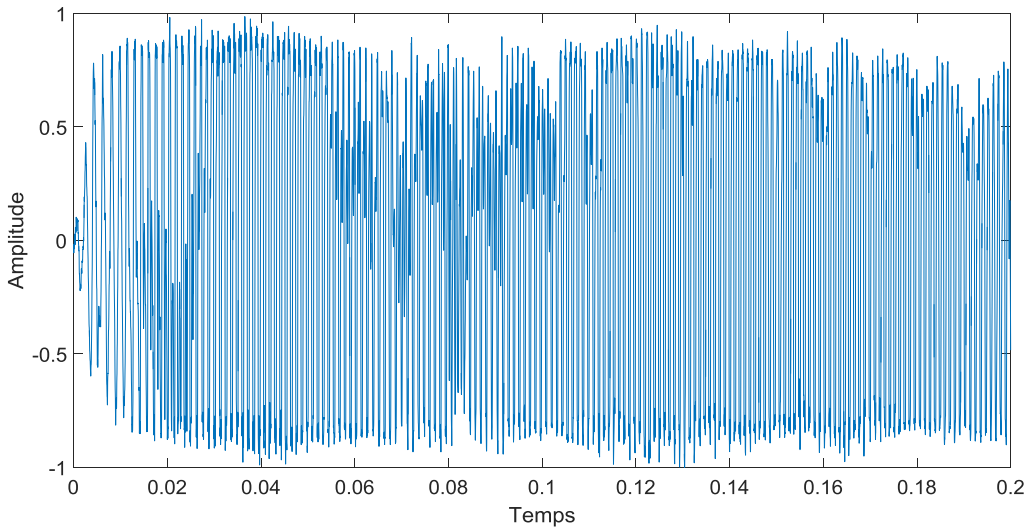

Figure 4.16 : Limitation de la durée d'un son de test à 0.2s

- **Etape 3 : Fixation du nombre d'attributs acoustiques**

Dans cet exemple, nous calculons 5 attributs MFCCs.

Les deux matrices suivantes montrent les valeurs numériques d'attributs calculés avant et après l'opération de normalisation. Le nombre de ligne correspond nombre de trames et le nombre de colonne correspond aux attributs.

1.0000	-2.5001	2.5973	-1.7264	0.8230	-0.1665
1.0000	-2.4572	2.2948	-1.1948	0.5157	-0.1322
1.0000	-2.3179	2.0744	-1.2287	0.7029	-0.1889
1.0000	-2.7021	3.0435	-1.6910	0.3022	0.1099
1.0000	-1.9722	1.5638	-0.9705	0.6216	-0.1133
1.0000	-1.9871	1.5687	-1.1292	0.9089	-0.2694

Attributs avant normalisation

0	-0.7789	0.6392	-0.3142	-0.2903	0.6628
0	-0.5538	-0.1309	0.9638	-1.6391	1.6060
0	-0.7851	0.0621	1.1758	-2.1876	2.0649
0	-0.9966	0.5829	0.4632	-1.5188	1.6209
0	-1.0681	0.9237	-0.2348	-0.6492	0.9223
0	-1.0576	1.0083	-0.5224	-0.1889	0.4735

Attributs après normalisation

La matrice M correspond aux données normalisées. Vu que le premier élément de cette matrice est une constante, il n'a pas été utilisé pour la constitution des séquences.

- **Etape 4 : calcul de la distance DTW entre la matrice M et le triplet T(i) pour i=1 jusqu'à 30,**

A titre d'exemple, les distances entre la matrice M et les trois matrices du triplet (1) sont :

d1= 8.9337, d2=20.6486 et d3=18.6338.

La décision préliminaire Dec1 pour ce cas est la classe 1 (cris humain).

- **Etape 5 : Constitution du vecteur des décisions finales {Dec1, Dec2, …. Dec30}**

Cette étape calcul les décisions obtenues en utilisant les 30 triplets du Codebook (Figure 4.17).

(1) cris humains,

(2) bris de glace,

(3) alarmes de voitures,

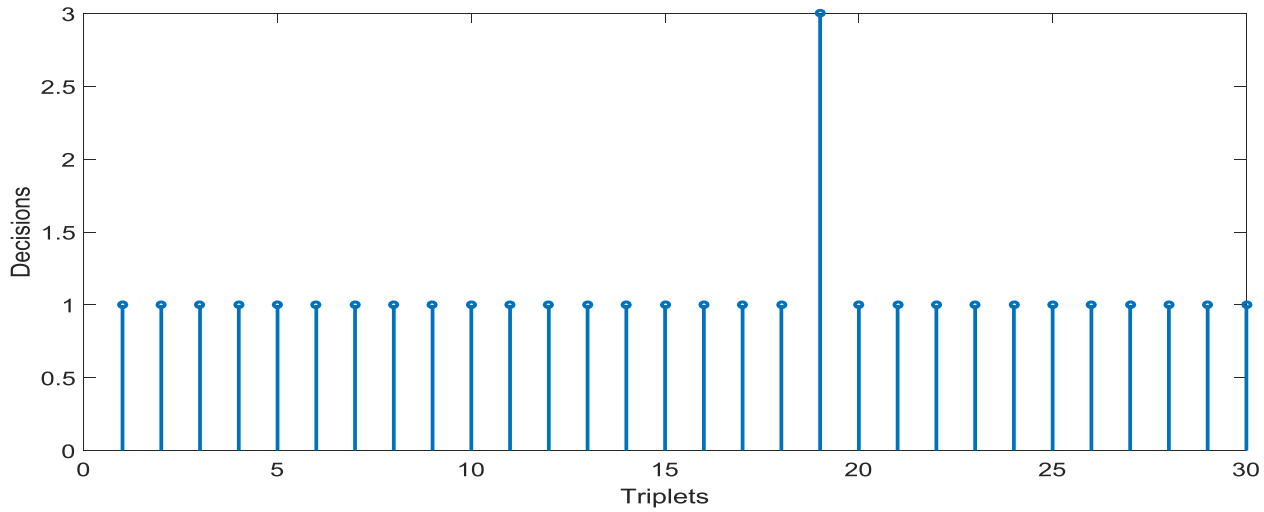

Figure 4.17 : Calcul des décisions finales

- **Etape 6 : Identification de la classe la plus probable**

D'après la Figure 4.17, nous remarquons que la classe la plus fréquente qui correspond au mode de cette suite est la classe 1 (cris humain).

4.4 Comparaison entre les LPCs et les MFCCs

Pour concevoir une méthode de reconnaissance à complexité réduite, nous comparons les résultats obtenus par les MFCCs et les LPCs en tenant compte de la contrainte cité ci-dessous :

- Trouver la durée minimale des sons de test qui aboutissent à un taux de reconnaissance élevé en utilisant le moins d'attributs possible,

La durée est variée entre 0.1 s et 0.6 s. Le nombre de coefficients est varié de 2 à 22.

4.4.3 LPCs

La Figure 4.18 présente les taux de classification (Acc) obtenus en variant le nombre des LPCs et la durée des sons.

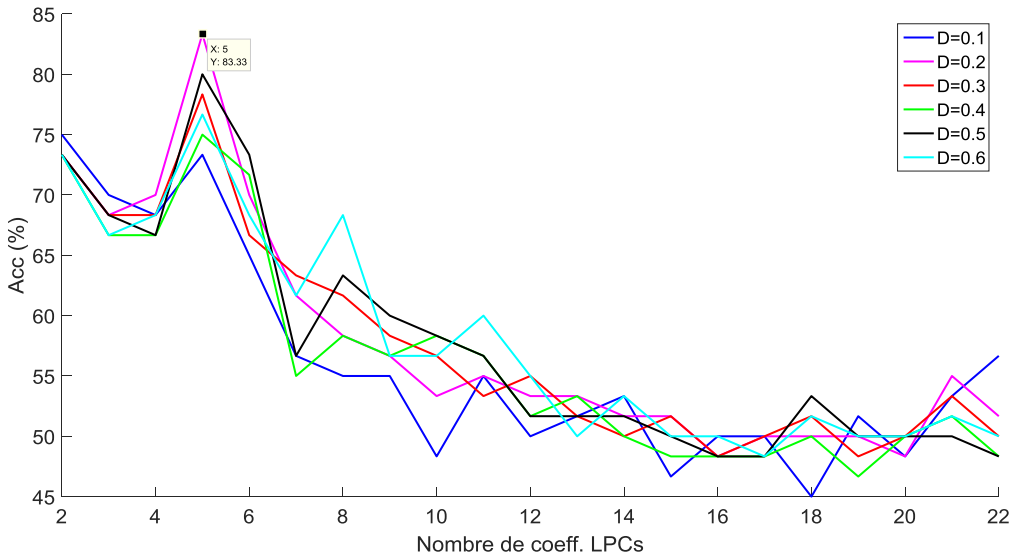

Figure 4.18 : Variation du nombre d'attributs LPCs et de la durée des sons de test

D'après la Figure 4.18, nous remarquons que pour une durée de 0.2s, l'Acc a atteint une valeur maximale de **83.33%** en utilisant seulement 5 coefficients LPCs. Le taux d'erreur correspondant est de : **16.66 %.**

Dans ce cas, la matrice de confusion est donnée comme suit (Figure 4.19):

classes prédites \ Classes réelles	A	B	C
A	17	0	6
B	0	19	0
C	3	1	14

Figure 4.19 : Matrice de confusion (cas des LPCs)

A, B et C sont les classes à reconnaitre :

A : Cris des humains,

B : Bris de glace,

C : Alarmes des voitures.

Les taux de précision et de rappel pour les trois classes sont résumés dans le tableau 4.1.

Tableau 4-1 : Précision et rappel du modèle de reconnaissance en utilisant 5 attributs LPCs

Métriques / Classes	Rappel (%)	Précision (%)
Cris humains	85	73.9
Bris de glace	95	100
Alarmes de voitures	70	77.7

A partir des résultats du tableau 4.1, nous avons obtenu des taux élevés du rappel et de la précision pour les trois classes. Nous pouvons conclure que les trois classes sont parfaitement gérées par cette configuration.

4.4.4 MFCCs (sans l'inclusion des dérivées)

La Figure 4.20 présente les Acc obtenus en variant le nombre des MFCCs et la durée des sons.

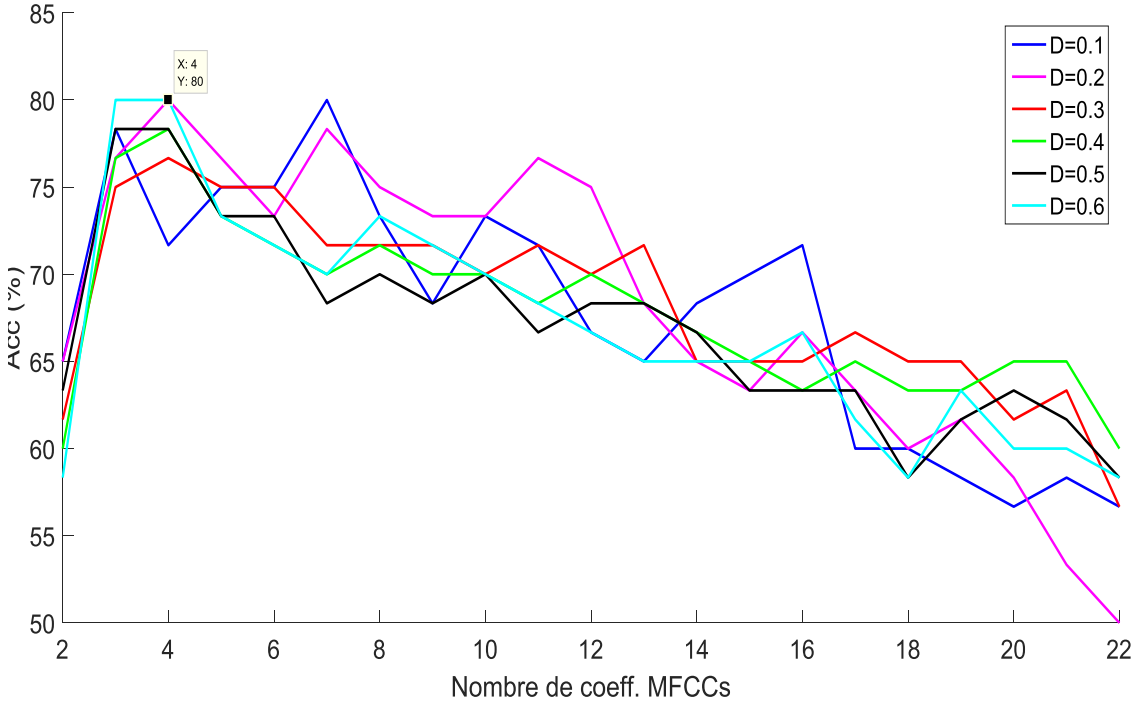

Figure 4.20 : Variation du nombre de MFCCs et de la durée des sons de test

D'après cette Figure, nous remarquons que pour une durée D=0.2s, l'Acc a atteint une valeur de 80% en utilisant seulement 4 coefficients MFCCs. Le taux d'erreur est de 20%.

Dans ce cas, la matrice de confusion est schématisée dans la Figure 4.21:

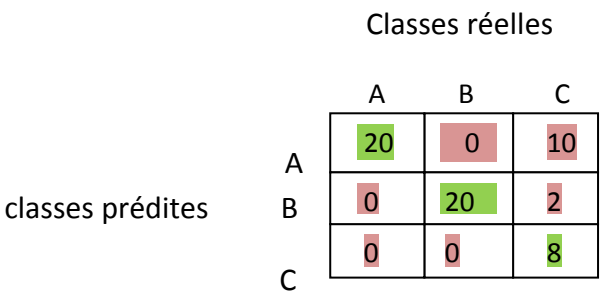

Figure 4.21 : Matrice de confusion (cas des MFCCs sans l'inclusion des dérivées)

Pour ce cas, les taux de précision et de rappel pour les trois classes sont résumés dans le tableau 4.2.

Tableau 4-2 : Précision et rappel du modèle de reconnaissance en utilisant 4 MFCCs

Métriques / **Classes**	**Rappel**	**Précision**
Cris humains	100%	66%
Bris de glace	100%	90.9%
Alarmes de voitures	40%	100%

A partir des résultats du tableau 4.2, nous remarquons :

- **Classe « Cris humains » :** le rappel est élevé mais la précision est faible : La classe est bien détectée mais le modèle comprend également des points d'autres classes.
- **Classe « Bris de glace » :** le rappel et la précision sont élevés : la classe est parfaitement gérée par cette configuration du modèle.
- **Classe « Alarmes de voitures » :** le rappel est faible mais la précision est haute : le modèle ne détecte pas correctement la classe, mais il est hautement fiable quand il le fait.

4.4.5 MFCCs (avec inclusion des deltas et delta delta)

La Figure (4.22) présente les résultats obtenus en incluant les drivées des MFCCs .

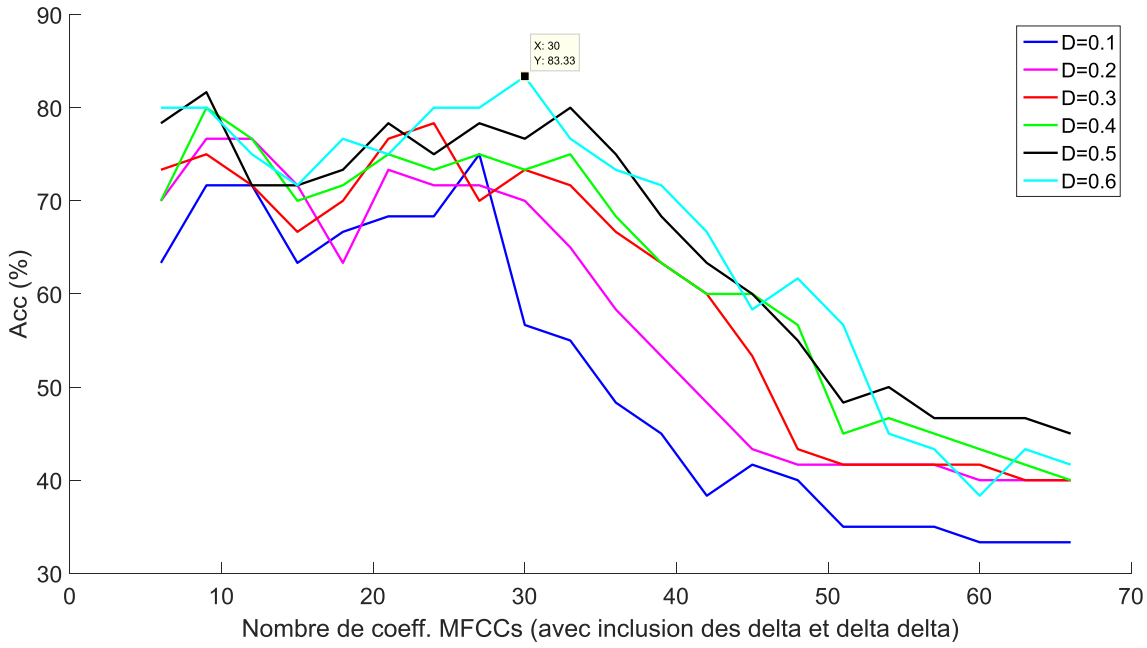

Figure 4.22 : Variation du nombre de MFCCs et de la durée des sons de test (avec inclusion des des drivées)

La matrice de confusion est donnée comme suit (Figure 4.23) :

		Classes réelles		
		A	B	C
classes prédites	A	15	0	0
	B	0	20	5
	C	5	0	15

Figure 4.23 : Matrice de confusion (cas des MFCCs avec inclusion des drivées)

D'après cette Figure, nous remarquons que pour une durée D=0.6s, l'Acc a atteint une valeur de 83,33% en utilisant 30 coefficients MFCCs. Le taux d'erreur est de 16.66%.

Tableau 4-3 : Précision et rappel du modèle de reconnaissance en utilisant 30 MFCCs (inclusion des drivées)

Métriques / Classes	**Rappel**	**Précision**
Cris humains	75%	100%
Bris de glace	100%	80%
Alarmes de voitures	75%	75%

A partir des résultats du tableau 4.3, nous avons obtenu des taux élevés du rappel et de la précision pour les trois classes. Nous pouvons conclure que les trois classes sont parfaitement gérées par ce modèle.

4.5 Conclusion

Dans ce chapitre, nous avons présenté les résultats expérimentaux de la reconnaissance des événements sonores. Au début, nous avons fait une présentation sur les logiciels et les langages de programmation utilisés. Par la suite, nous avons présenté les différentes étapes nécessaires pour l'implémentation de la méthode proposée. Nous avons utilisé le taux de bonne classification, la précision, et le rappel pour évaluer les performances de la méthode.

La dernière partie de ce chapitre a été consacrée à une comparaison entre deux techniques d'extraction d'attributs : les MFCCs et les LPCs. D'après les résultats obtenus, nous avons trouvé que la méthode proposée a atteint un taux de reconnaissance de 83.33% en utilisant seulement 5 attributs LPCs. De plus, ce résultat a été obtenu en exploitant seulement une durée égale à 0.2s du segment sonore. Cependant, pour les MFCCs, le même taux de reconnaissance a été atteint mais en utilisant (i) une durée du segment sonore de 0.6s et (ii) 30 attributs (incluant les dérivées des MFCCs). On peut conclure alors que les LPCs sont plus appropriés dans ce cas.

Notre méthode possède l'avantage d'être :

(i) Rapide, efficace et non dépendante d'une phase d'apprentissage,

(ii) Implémentable en temps réel,

Cependant, le seul inconvénient de cette méthode est résumé comme suit :

L'utilisation de la DTW pour des séquences de longues durées nécessite un temps de calcul considérable. Cependant, dans notre contexte et d'après les résultats obtenus, nous avons remarqué que la meilleure performance a été atteinte en utilisant une durée de 0.2s avec seulement 5 attributs. Les paramètres de notre méthode n'ont pas de grande influence sur le temps de calcul et la vitesse d'exécution.

Chapitre 5 : Conclusions et travaux futures

5.1 Conclusions

Ce travail consiste en la conception d'une méthode de reconnaissance des évènements sonores en vue de la mise en œuvre d'un système de surveillance audio. La méthode que nous proposons utilise peu de données pour classer les sons. Elle est basée principalement sur la déformation temporelle dynamique.

Nous avons présenté dans le deuxième chapitre du mémoire les généralités sur les systèmes de surveillance. Nous nous sommes focalisés sur l'apport de la modalité audio dans la surveillance des lieux.

Dans le troisième chapitre, nous avons présente la méthode de reconnaissance. Notre objectif dans cette étude est de comparer les performances de la reconnaissance en utilisant deux techniques d'extraction d'attributs : les MFCCs et les LPCs. Les MFCCs sont basés sur le processus de la perception auditive humaine tandis que les LPCs sont estimées à partir du modèle phonatoire de production de la parole. Nous avons utilisé une multitude de données de références (codebook) pour le calcul des distances temporelles entre les séquences. Nous avons identifié la classe la classe la plus fréquence dans notre schéma de classification en utilisant le mode.

Nous nous sommes limités dans notre projet à trois catégories de sons : (i) cris humains, (ii) alarmes de voitures et (iii) bris de glace. La méthode proposée a été présentée en détails dans le troisième chapitre. Elle tient compte des exigences suivantes :

- La reconnaissance des sons de l'environnement doit utiliser des méthodes à complexité réduite pour faciliter son implémentation en temps réel,
- Le module de reconnaissance doit être adapté à une plateforme de détection des sons impulsifs qui existe au niveau du laboratoire ASM,
- Le début exact du son est considéré connu,
- La durée du son nécessaire à la tâche de reconnaissance doit être courte,

Dans le quatrième chapitre, nous avons présenté les résultats expérimentaux. Nous avons effectué une comparaison entre les MFCCs et les LPCs. Nous avons utilisé le taux de bonne classification, la précision, et le rappel pour évaluer les performances de la méthode proposée.

D’après les résultats obtenus, nous avons trouvé que la méthode proposée a atteint un taux de reconnaissance de 83.33**%** en utilisant seulement 5 attributs LPCs.

De plus, ce résultat a été obtenu en exploitant seulement une durée égale à 0.2s du segment sonore. Cependant, pour les MFCCs, le même taux de reconnaissance a été atteint mais en utilisant (i) une durée du segment sonore de 0.6s et (ii) 30 attributs (incluant les dérivées des MFCCs). Nous concluons que les LPCs sont plus appropriés dans ce cas.

Notre méthode possède l’avantage d’être :

(i) Rapide, efficace et non dépendante d’une phase d’apprentissage,

(ii) Implémentable en temps réel,

5.2 Travaux futurs

Les travaux futurs consistent à :

- Augmenter le nombre de classes et mesurer les performances de la méthode,
- Intégrer ce module de reconnaissance dans la plateforme de détection des sons impulsifs en temps réel,
- Mesurer les performances de détection et reconnaissance des sons de l’environnement,
- Comparer le système proposé avec les solutions existantes.

Bibliographie

[1] M. A. Najjar, M. Ghantous and M. A. Bayoumi, Video Surveillance for Sensor Platforms - Algorithms and Architectures, vol. 175, Lecture Notes in Electrical Engineering, 2014, pp. 1-202.

[2] W. Hu, T. Tan, L. Wang and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews,* vol. 34, no. 3, pp. 334-352, 2004.

[3] R. T. Collins, A. J. Lipton and T. Kanade, "Introduction to the special section on video surveillance," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 22, no. 2, pp. 745-746, 2000.

[4] H. N. Tipton, «Information security management handbook.,» sixth edition,Taylor et francis, 2007.

[5] M. B. Brown, "Home security system utilizing television surveillance," U.S. Patent, No 3.482,037, 1969.

[6] S.Lecomte, "Classification partiellement supervisée par SVM application à la détection d'evenements en surveillance audio," Thése de Doctorat, Université de Technologie de Troyes, France, 2013.

[7] Y.-K. Ki and D.-K. Baik, "Model for accurate speed measurement using double-loop," *IEEE Transactions on Vehicular Technology,* vol. 55, no. 4, pp. 1094-1101, 2006.

[8] C. Micheloni, G. L. Foresti and L. Snidaro, "A co-operative multicamera system for videosurveillance of parking lots," in *IEEE Symposium on Intelligent Distributed Surveillance Systems*, London, 2003.

[9] G. Barrenetxea, F. Ingelrest, G. Schaefer and M. Vetterli, "Wireless sensor networks for environmental monitoring: the sensor scope experience," in *IEEE International Zurich Seminar on Communications*, Zurich, 2008.

[10] L. Cutrona, W. Vivian, E. Leith and G. Hall, "A high-resolution radar combat-surveillance system," *IRE Transactions on Military Electronics,* Vols. MIL-5, no. 2, pp. 127-131, 2009.

[11] J. Wang, C. Qimei, Z. De and B. Houjie, "Embedded wireless video surveillance system for vehicle," in *International Conference on Telecommunications*, Chengdu, China, 2006.

[12] S. Fleck and W. Strasser, "Smart camera-based monitoring system and its application to assisted living," *Proceedings of the IEEE,* vol. 96, no. 10, pp. 1698-1714, 2008.

[13] C. P. Diehl, Toward efficient collaborative classification for distributed video surveillance, Pittsburgh, 2000.

[14] Z. Zhu and T. S. Huang, Multimodal surveillance: sensors, algorithms, and systems, Artech House, 2007.

[15] F. Ykhlef, S. A. Hamada, F. Ykhlef, A. Derbal and D. Bouchaffra, "Real-Time detection of impulsive sounds for audio surveillance systems," in *The national study day on research on computer sciences*, Saida, Algeria, 2019.

[16] M.Aramki, "Analyse-Synthése de Sons Impulsifs:Approches Physique et Perceptive," Thése de Doctorat, Université de la Méditerranee-Aix-Marseille, 2003.

[17] A.Dufaux, "Detection and recognition of impulsive sound signals," Phd Thesis,Neuchatel University, Switzerland, 2001.

[18] C.Rosin, B. B and B.Defreville, "Monitoring du Bruit des Avions: Une Detection à Partir du signal audio," in *Congrés Français d'acoustique*, Lyon,France., 2010.

[19] W. Ma, H. Bao, C. Zhang and X. Liu, "Beamforming of phased microphone array for rotating sound source localization," *Journal of Sound and Vibration,* vol. 447, p. 115064, 2020.

[20] S.Souli and Z.Lachiri, "Audio classification using scattering features and support vectors machines for ledical surveillance," *Applied Acoustics,* vol. 130, pp. 270-282, 2018.

[21] A.Rabaoui, M.Davy, S.Rossignol, Z.Lachiri and N.Ellouze, "Selection de descripteurs audio pour la classification des sons environnementaux avec des SVMs mono-classe," in *Colloque GRETSI*, France, 2007.

[22] S. Rovetta, Z. Mnasri and F. Masulli, "Detection of hazardous road events from audio streams: An ensemble outlier detection approach," in *IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, Bari, Italy, 2020.

[23] P. Gheorghe, A. Caranica, H. Cucu and Burileanu, "Sound event recognition in smart environments," in *IEEE International Conference on Speech Technology and Human-Computer Dialogue*, Romania, 2015.

[24] H.-R. Choi and K. TaeYong, "Modified dynamic time warping based on direction similarity for fast gesture recognition," *Mathematical Problems in Engineering,* vol. 2018, 2018.

[25] L. Calliope and G. Fant, La parole et son traitement automatique, Paris: Masson, 1989.

[26] "Youtube," [Online]. Available: www.youtube.com. [Accessed March 2020].

[27] "Sounddogs," [Online]. Available: https://www.sounddogs.com/. [Accessed March 2020].

[28] P. Pere, M. Dušan and N. Climent, "On real-time mean-and-variance normalization of speech recognition features," in *IEEE International conference on acoustics speech and signal processing*, Barcelona, Spain, 2006.

[29] N. Dave, "Feature extraction methods LPC, PLP and MFCC in speech recognition," *International Journal for Advance Research in Engineering and Technology,* vol. 1, no. 6, pp. 1-4, 2013.

[30] S. URMILA, «TECHNIQUES FOR FEATURE EXTRACTION IN SPEECH».

[31] M. Rishiraj, "Speaker Recognition Using Shifted MFCC," Thése de Master, University of South Florida, Florida, USA, 2012.

[32] A. Elkourd, Arabic isolated word speaker dependent recognition system, LAP LAMBERT Academic Publishing, 2014, pp. 26-51.

[33] A. Sithara, A. Thomas and D. Mathew, "Study of MFCC and IHC feature extraction methods with probabilistic acoustic models for speaker biometric applications," *Procedia computer science,* vol. 143, pp. 267-276, 2015.

[34] J. Bradbury, Linear Predictive Coding, Mc G.Hill, 2000.

[35] J. Hai and E. M. Joo, "Improved Linear Predictive Coding Method for Speech Recognition.," in *IEEE. Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia*, 2003.

[36] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 23, no. 1, pp. 67-72, 1975.

[37] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing ,* vol. 26, no. 1, pp. 43-49, 1978.

[38] B.-X. Chen, "A fast algorithm for dynamic time warping with adaptive window," Master thesis, National Sun Yat-sen University, 2018.

[39] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics,* pp. 1-14, 2018.

[40] "Towards data science," [Online]. Available: https://towardsdatascience.com/. [Accessed October 2020].

[41] "MATLAB," [Online]. Available: https://fr.mathworks.com/. [Accessed March 2020].

[42] "Voicebox," [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html. [Accessed October 2020].

[43] Q. Wang, "DTW," [Online]. Available: https://wangquan.me. [Accessed October 2020].

[44] "GoldWave," [Online]. Available: https://www.goldwave.com/. [Accessed March 2020].

[45] S. A. Hamada, "Détection des sons impulsives en vue de la mise en oeuvre d'un systéme de surveillance audio," Mémoire de Master, Université SAAD DAHLEB, Blida, Algérie, 2017.